All Training materials are provided "as is" and without warranty and RStudio disclaims any and all express and implied warranties including without limitation the implied warranties of title, fitness for a particular purpose, merchantability and noninfringement.

Studio

Tidy data

Prepare data faster with reshape2

https://www.flickr.com/photos/jamesgibbard/4300994347

Garrett Grolemund

Master Instructor, RStudio

August 2014

1. Loading data
2. Reformatting data / Tidy data
3. Saving data

Tidy data

What is tidy data?

- Data that is easy to model, visualise and aggregate (i.e. works well with `lm`, `ggplot`, and `dplyr`)
- A step along the road to clean data

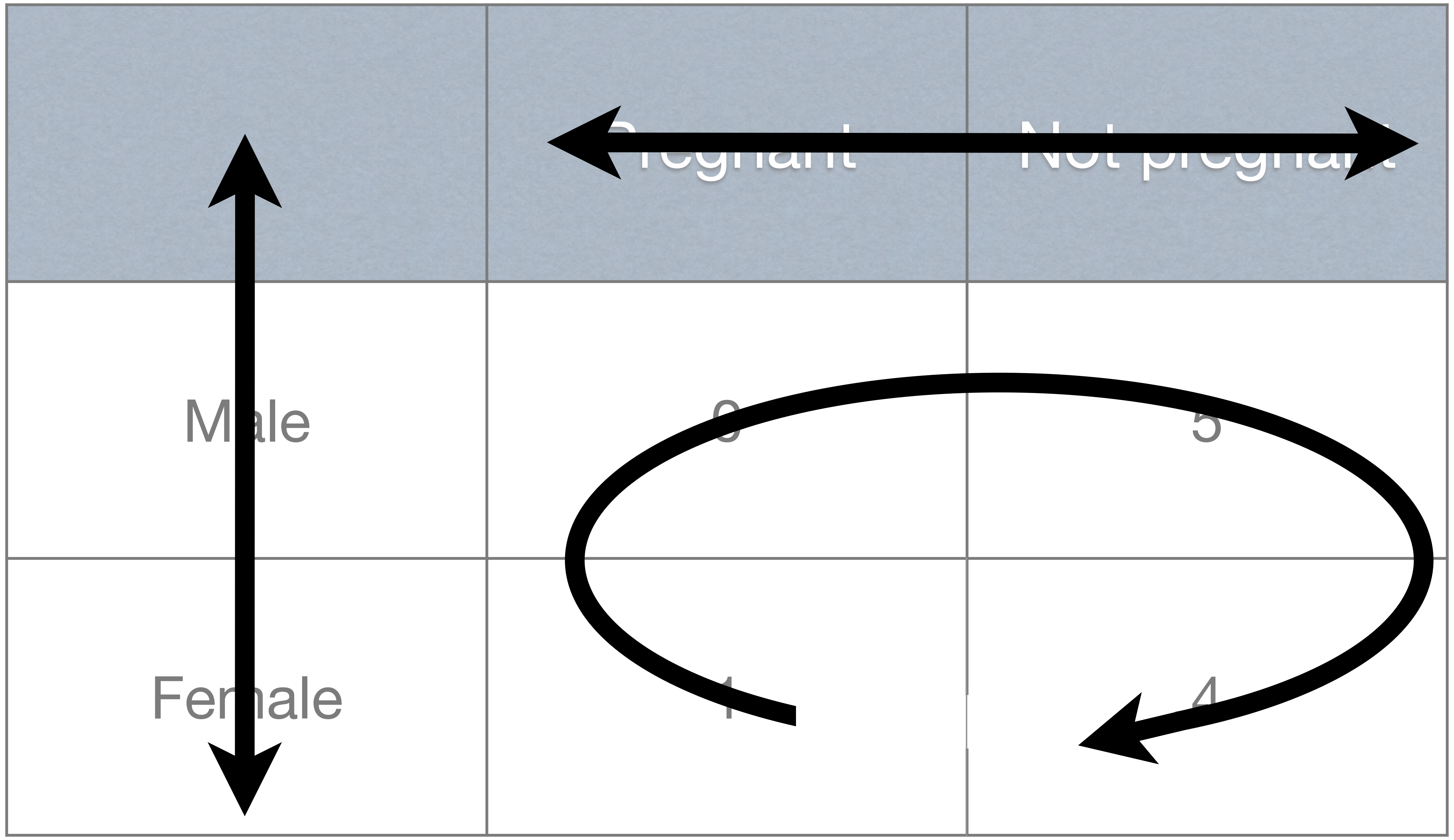
Tidy Data

Storage	Contains
Rows	Observations
Columns	Variables
One data frame	Entire data se

Many useful R functions
expect tidy data!

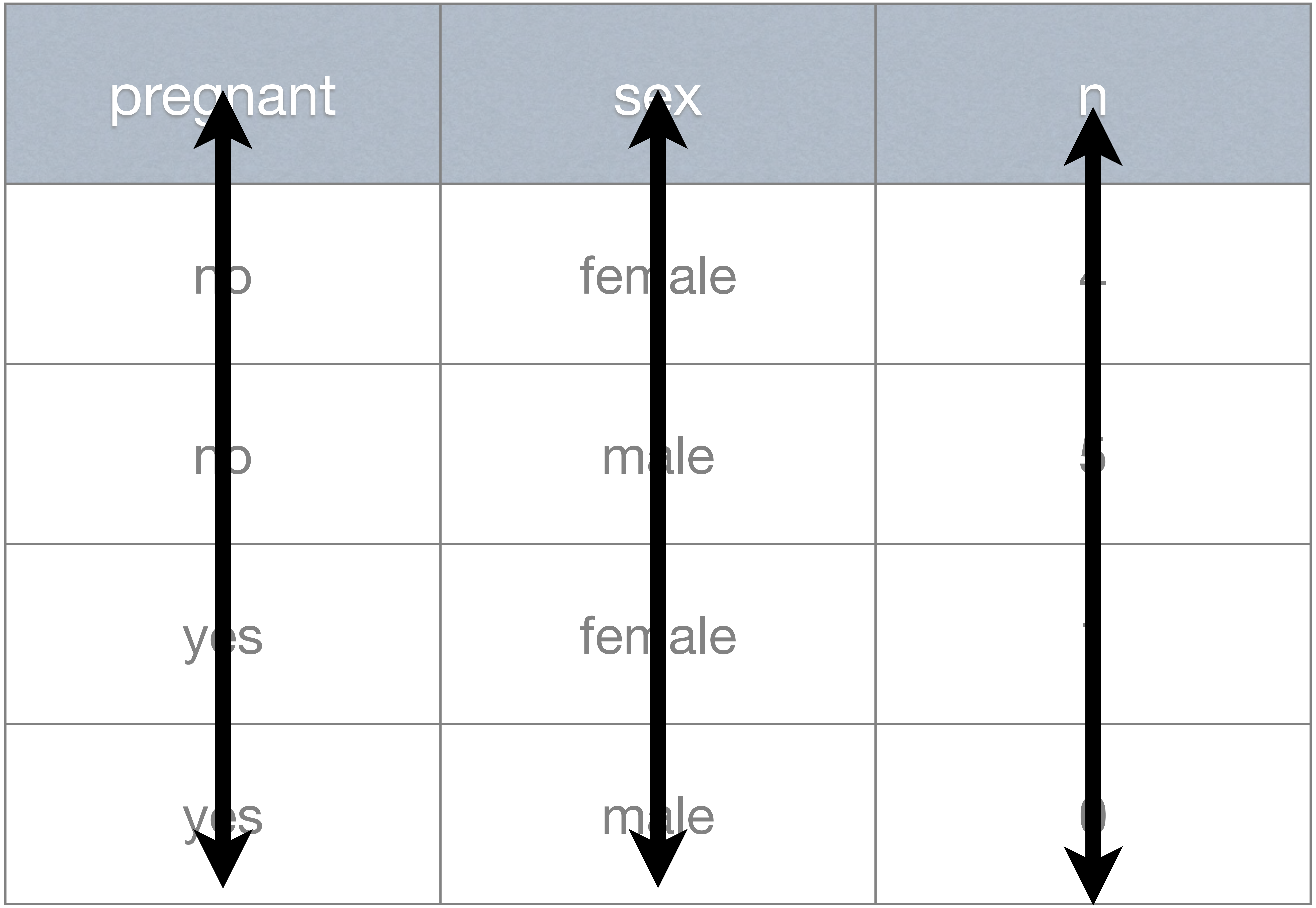
	Pregnant	Not pregnant
Male	0	5
Female	1	4

There are three variables in this data set.
What are they?



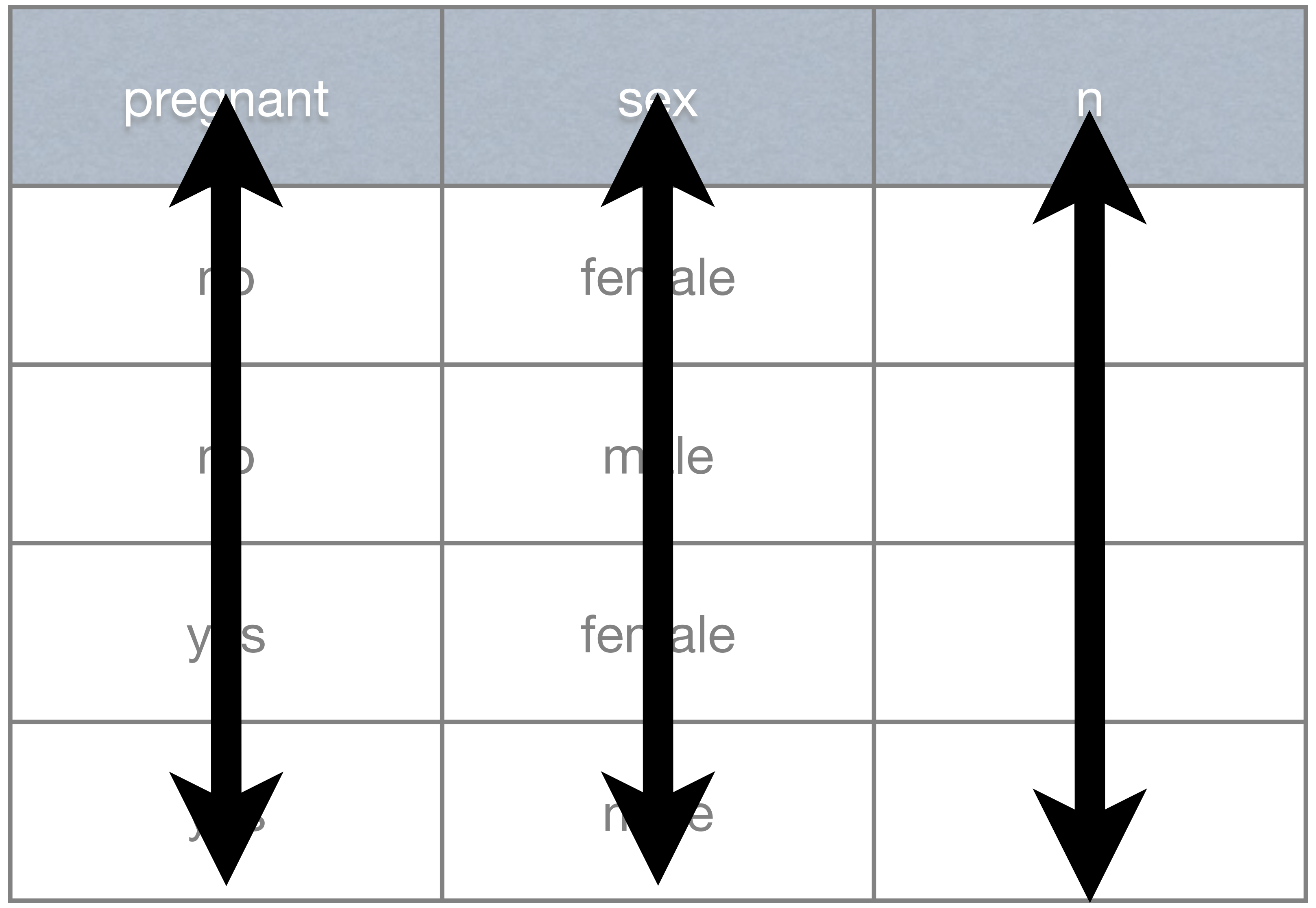
	Pregnant	Not pregnant
Male	0	5
Female	1	4

pregnant	sex	n
no	female	4
no	male	5
yes	female	1
yes	male	0

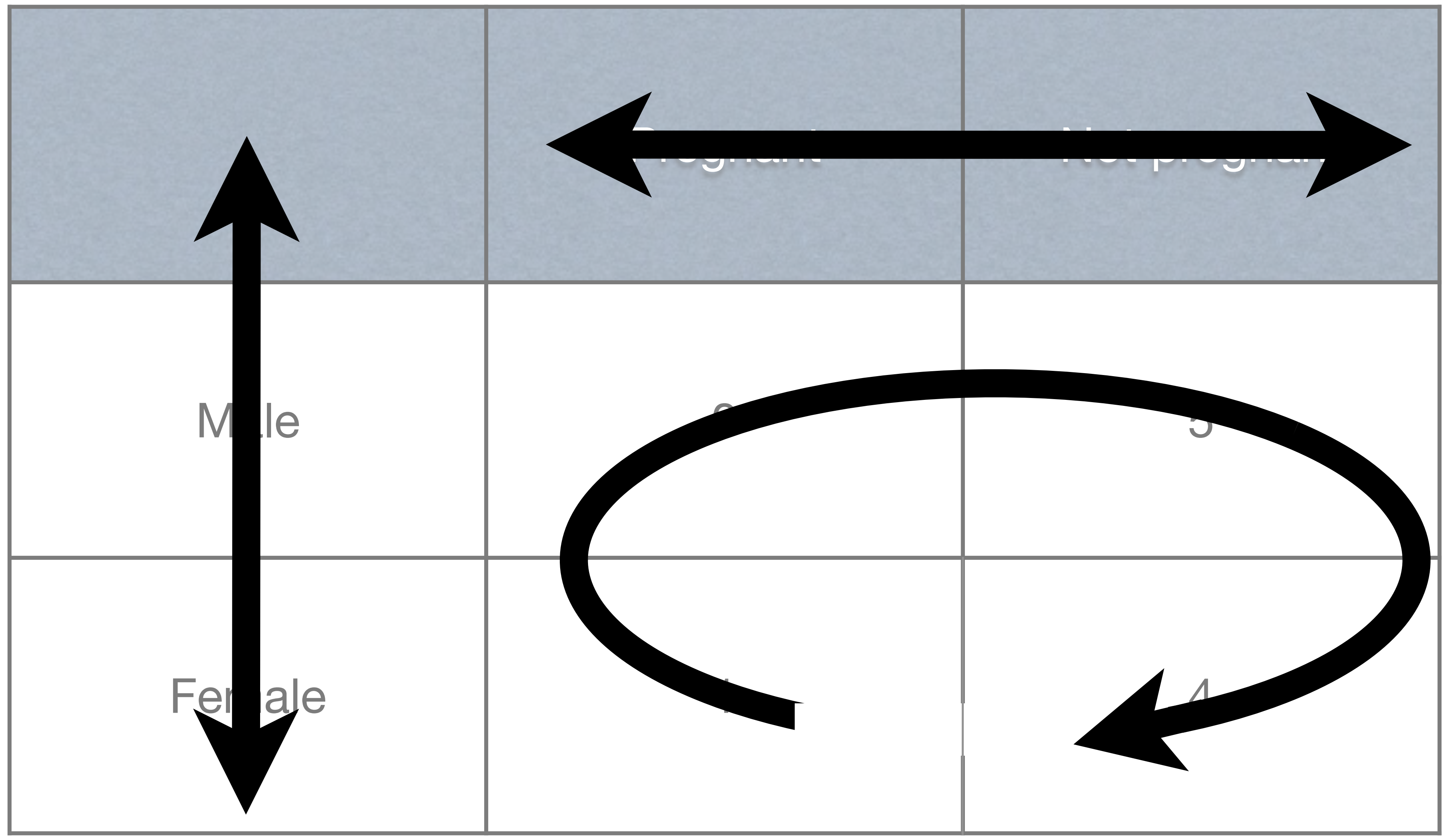


pregnant	sex	n
no	female	4
no	male	5
yes	female	1
yes	male	0

Which would you rather work with?



pregnant	sex	n
no	female	
no	male	
yes	female	
yes	male	



	Pregnant	Not pregnant
Male	0	5
Female	1	4

```
df$pregnant
df$sex
df$n
```

```
df[[1]]
names(df)
c(df[2,2],df[3,2],df[2,3],df[3,3])
```

Common causes of messiness

- column headers are values, not variable names
- cells are variable names, not values
- data split over multiple files

Values in column names

Income distribution within U.S. religious groups

CC BY http://www.flickr.com/photos/52890443@N02/4890869149

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

Loading data

Make sure the file is in your working directory.

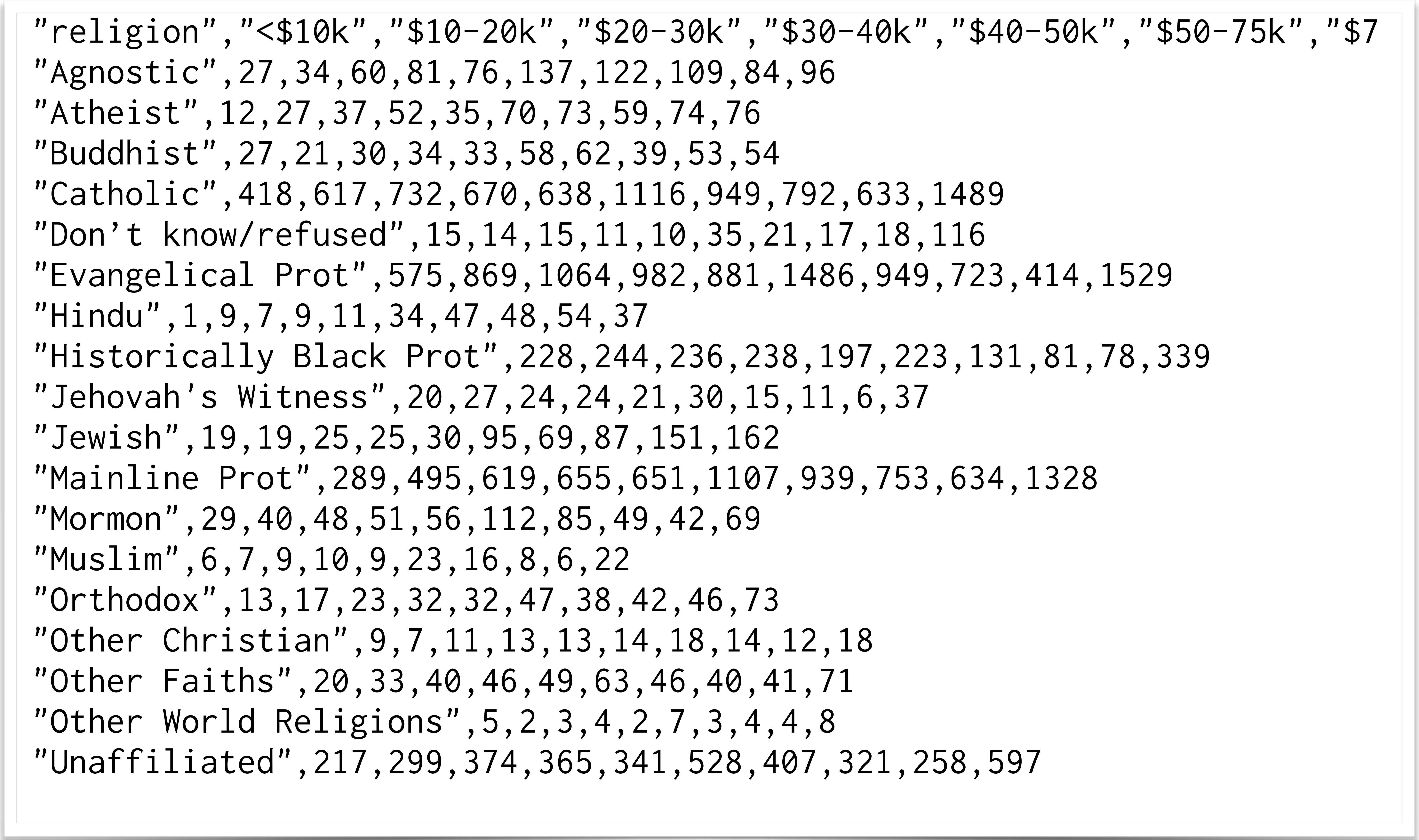
```
raw <- read.csv(”data/pew.csv”, check.names = F)
```

Name of file
to read

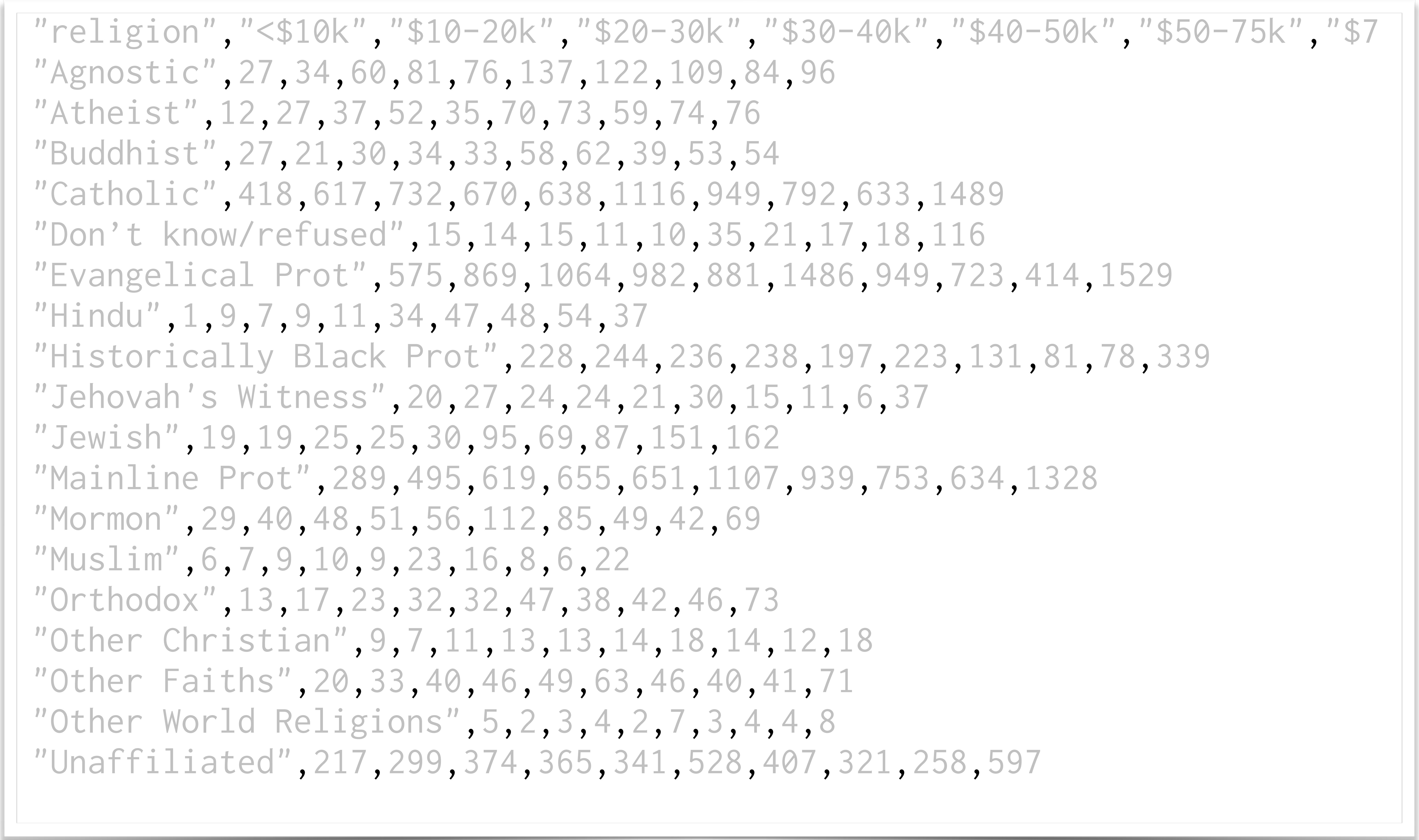
```
raw <- read.csv(”data/pew.csv”, check.names = F)
```

read function, based on file's separator character

```
raw <- read.csv(”data/pew.csv”, check.names = F)
```



```
"religion","<$10k","$10-20k","$20-30k","$30-40k","$40-50k","$50-75k","$7
"Agnostic",27,34,60,81,76,137,122,109,84,96
"Atheist",12,27,37,52,35,70,73,59,74,76
"Buddhist",27,21,30,34,33,58,62,39,53,54
"Catholic",418,617,732,670,638,1116,949,792,633,1489
"Don’t know/refused",15,14,15,11,10,35,21,17,18,116
"Evangelical Prot",575,869,1064,982,881,1486,949,723,414,1529
"Hindu",1,9,7,9,11,34,47,48,54,37
"Historically Black Prot",228,244,236,238,197,223,131,81,78,339
"Jehovah's Witness",20,27,24,24,21,30,15,11,6,37
"Jewish",19,19,25,25,30,95,69,87,151,162
"Mainline Prot",289,495,619,655,651,1107,939,753,634,1328
"Mormon",29,40,48,51,56,112,85,49,42,69
"Muslim",6,7,9,10,9,23,16,8,6,22
"Orthodox",13,17,23,32,32,47,38,42,46,73
"Other Christian",9,7,11,13,13,14,18,14,12,18
"Other Faiths",20,33,40,46,49,63,46,40,41,71
"Other World Religions",5,2,3,4,2,7,3,4,4,8
"Unaffiliated",217,299,374,365,341,528,407,321,258,597
```



```
”religion”,”<$10k”,”$10-20k”,”$20-30k”,”$30-40k”,”$40-50k”,”$50-75k”,”$7
”Agnostic”,27,34,60,81,76,137,122,109,84,96
”Atheist”,12,27,37,52,35,70,73,59,74,76
”Buddhist”,27,21,30,34,33,58,62,39,53,54
”Catholic”,418,617,732,670,638,1116,949,792,633,1489
”Don’t know/refused”,15,14,15,11,10,35,21,17,18,116
”Evangelical Prot”,575,869,1064,982,881,1486,949,723,414,1529
”Hindu”,1,9,7,9,11,34,47,48,54,37
”Historically Black Prot”,228,244,236,238,197,223,131,81,78,339
”Jehovah’s Witness”,20,27,24,24,21,30,15,11,6,37
”Jewish”,19,19,25,25,30,95,69,87,151,162
”Mainline Prot”,289,495,619,655,651,1107,939,753,634,1328
”Mormon”,29,40,48,51,56,112,85,49,42,69
”Muslim”,6,7,9,10,9,23,16,8,6,22
”Orthodox”,13,17,23,32,32,47,38,42,46,73
”Other Christian”,9,7,11,13,13,14,18,14,12,18
”Other Faiths”,20,33,40,46,49,63,46,40,41,71
”Other World Religions”,5,2,3,4,2,7,3,4,4,8
”Unaffiliated”,217,299,374,365,341,528,407,321,258,597
```

`read.csv()`: comma separated

`read.delim()`: tab separated

`read.delim(sep = ”|”)`: | separated

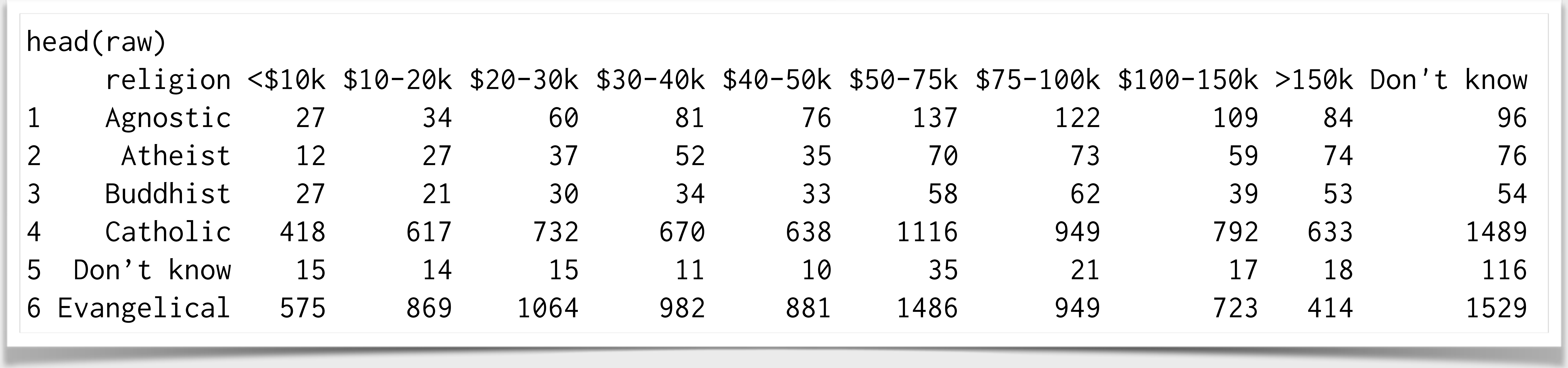
`read.fwf()`: fixed width

```r
raw <- read.csv(”data/pew.csv”, check.names = F)
```

Not important.
The variable names in this data set begin with "$", which R would change to avoid possible problems. I'm telling R not to.

Your turn

What are the variables in this data set?

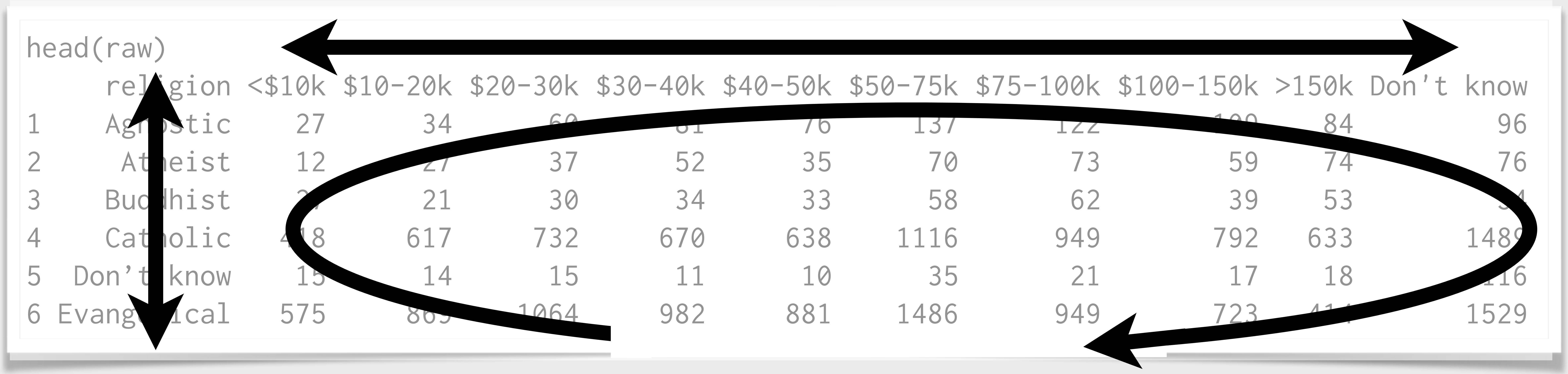


```
head(raw)
     religion <$10k $10-20k $20-30k $30-40k $40-50k $50-75k $75-100k $100-150k >150k Don’t know
1    Agnostic    27      34      60      81      76     137      122       109    84         96
2     Atheist    12      27      37      52      35      70       73        59    74         76
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
```

01:00

Your turn

What are the variables in this data set?



```
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
```

```
# Fixing this problem is easy.  We use melt, from
# reshape2, with two arguments, the input data, and
# the columns which are already variables:

library(reshape2)
tidy <- melt(raw, id = ”religion”)

head(tidy)
```

Melting data

```
tidy <- melt(raw, id = ”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
```

Melting data

data set to melt

```
tidy <- melt(raw, id = ”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
```

Melting data

data set to melt

column(s) to keep as is

```
tidy <- melt(raw, id = "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
```

Melting data

data set to melt

column(s) to keep as is

```
tidy <- melt(raw, id = "religion")
```

remaining columns are “melted” into 2 columns: variable and value

```
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
```

Melting data

data set to melt

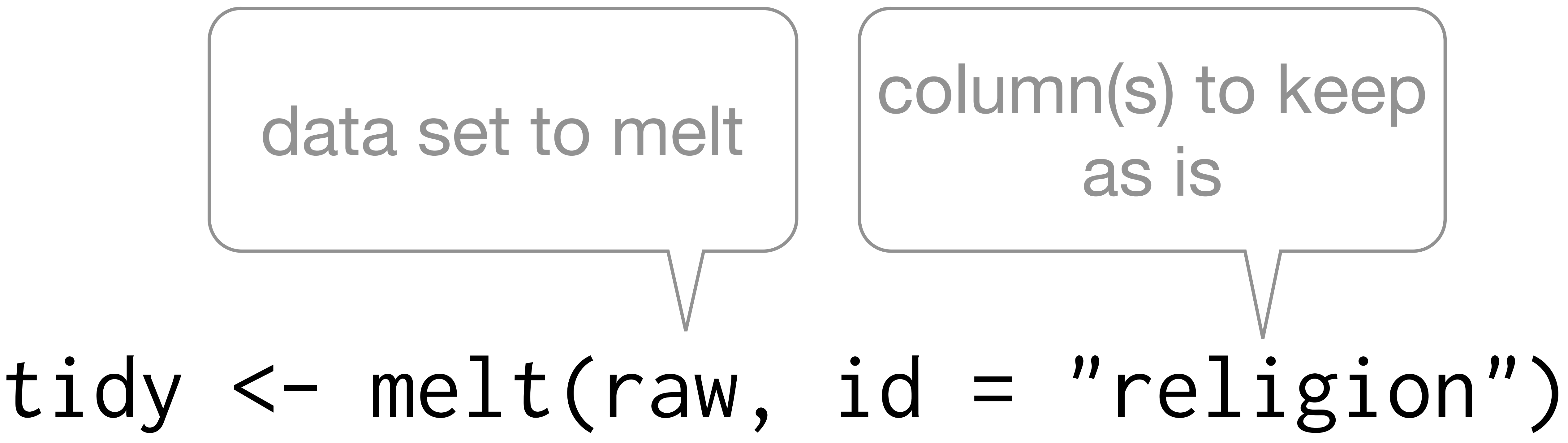
column(s) to keep as is

```
tidy <- melt(raw, id = ”religion”)
```

Column names are placed into one column, named "variable"

```
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
```

Melting data



data set to melt

column(s) to keep as is

```
tidy <- melt(raw, id = "religion")
```

Column names are placed into one column, named "variable"

```
head(tidy)
     religion variable    -
1    Agnostic    <$10k    -
2     Atheist    <$10k    -
3    Buddhist    <$10k    -
4    Catholic    <$10k    -
5  Don’t know    <$10k    -
6 Evangelical    <$10k    -
```

Melting data

data set to melt

column(s) to keep as is

```
tidy <- melt(raw, id = ”religion”)
```

Cell values are placed into a second column named "value"

```
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
```

Melting data

data set to melt

column(s) to keep as is

```
tidy <- melt(raw, id = ”religion”)
```

Cell values are placed into a second column named "value"

```
head(tidy)
     religion variable value
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
```

```
head(raw)
      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
```

```
head(tidy)
      religion variable value
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

```
head(raw)
      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
```

```
head(tidy)
      religion variable value
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

```
head(raw)
     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
```

```
head(tidy)
     religion variable value
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

```
head(raw)
     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
```

```
head(tidy)
     religion variable value
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

```
head(raw)
      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
```

```
head(tidy)
      religion variable value
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

```
head(raw)
     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
```

```
head(tidy)
     religion variable value
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

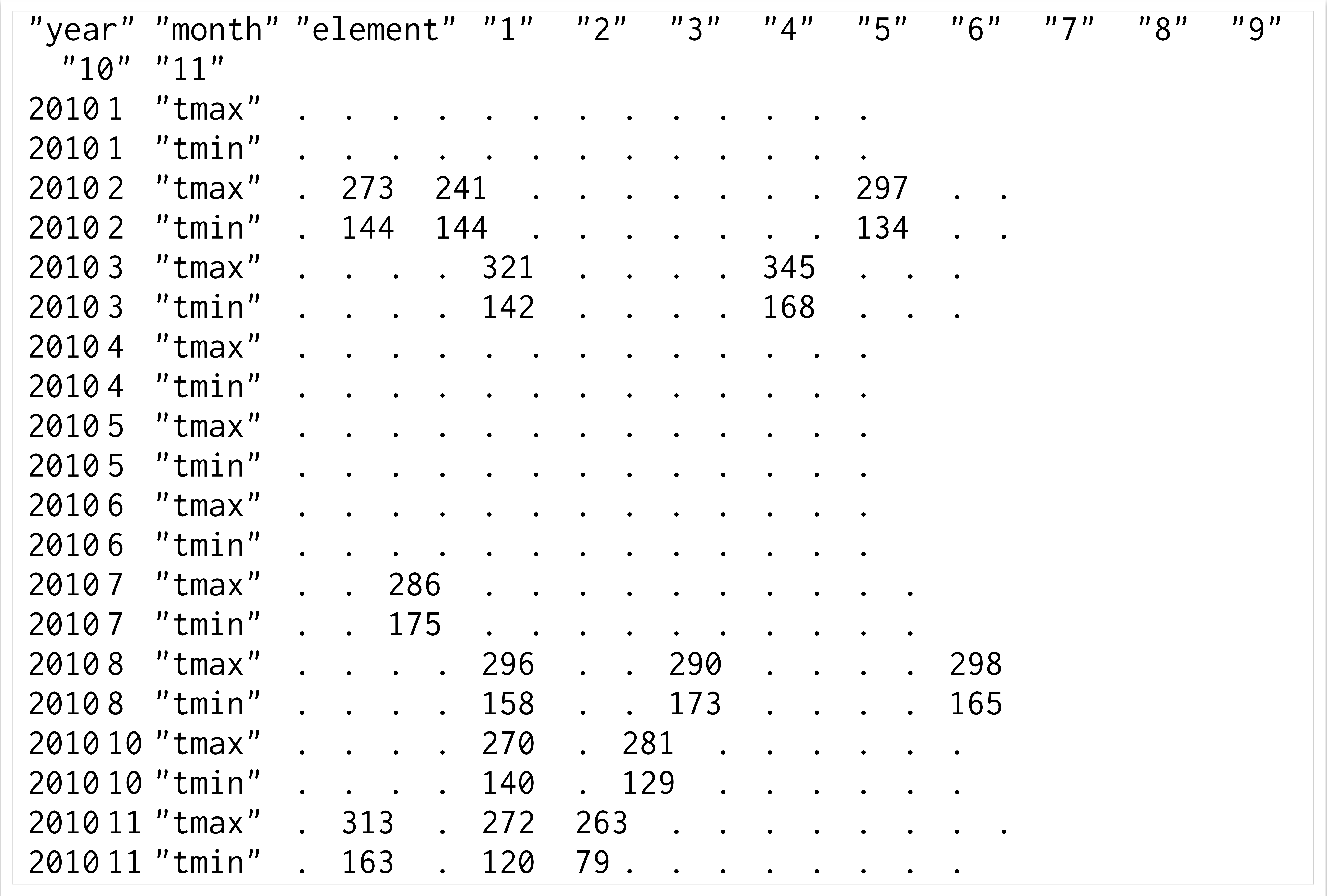
```
# We can now fix the column names
names(tidy) <- c("religion", "income", "n")

# Alternatively
tidy <- melt(raw, id = "religion",
  variable.name = "income", value.name = "n")
```

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



```
"year" "month" "element" "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"
  "10" "11"
2010 1  "tmax"  . . . . . . . . . . . . .
2010 1  "tmin"  . . . . . . . . . . . . .
2010 2  "tmax"  . 273 241 . . . . . . . 297 . .
2010 2  "tmin"  . 144 144 . . . . . . . 134 . .
2010 3  "tmax"  . . . . 321 . . . . 345 . . .
2010 3  "tmin"  . . . . 142 . . . . 168 . . .
2010 4  "tmax"  . . . . . . . . . . . . .
2010 4  "tmin"  . . . . . . . . . . . . .
2010 5  "tmax"  . . . . . . . . . . . . .
2010 5  "tmin"  . . . . . . . . . . . . .
2010 6  "tmax"  . . . . . . . . . . . . .
2010 6  "tmin"  . . . . . . . . . . . . .
2010 7  "tmax"  . . 286 . . . . . . . . . .
2010 7  "tmin"  . . 175 . . . . . . . . . .
2010 8  "tmax"  . . . . 296 . . 290 . . . . 298
2010 8  "tmin"  . . . . 158 . . 173 . . . . 165
2010 10 "tmax"  . . . . 270 . 281 . . . . . .
2010 10 "tmin"  . . . . 140 . 129 . . . . . .
2010 11 "tmax"  . 313 . 272 263 . . . . . . . .
2010 11 "tmin"  . 163 . 120 79 . . . . . . . .
```

```
raw <- read.delim(”data/weather.txt”,
  check.names = F, na.strings = ”.”)
```

```
raw <- read.delim(”data/weather.txt”,
  check.names = F, na.strings = ”.”)
```

Converts every . to an NA

Your turn

Melt the data to fix the days variable.

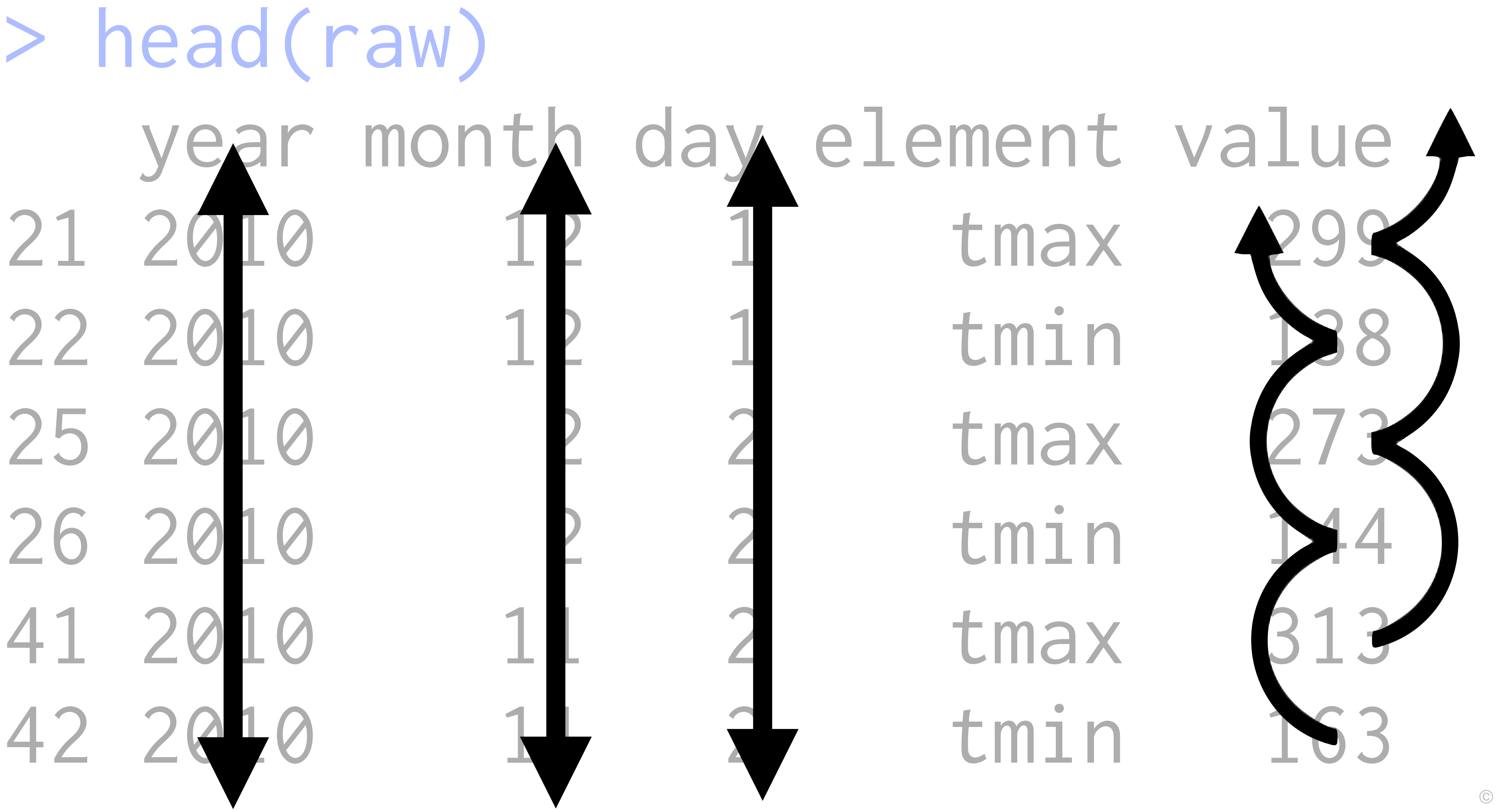
What do you need to do next?

```
# na.rm = TRUE is useful if the missing values don't have
# any meaning
raw <- melt(raw,
  id = c(”year”, ”month”, ”element”),
  variable.name = ”day”, na.rm = TRUE)

# reordering columns
raw <- raw[, c(”year”, ”month”, ”day”,
  ”element”, ”value”)]
```

```
# What are the variables in this dataset?
# Hint: tmin = minimum temperature
```

```
> head(raw)
   year month day element value
21 2010    12   1    tmax   299
22 2010    12   1    tmin   138
25 2010     2   2    tmax   273
26 2010     2   2    tmin   144
41 2010    11   2    tmax   313
42 2010    11   2    tmin   163
```

```
# What are the variables in this dataset?
# Hint: tmin = minimum temperature
```



```
> head(raw)
    year month day element value
21  2010    12   1    tmax   299
22  2010    12   1    tmin   138
25  2010     2   2    tmax   273
26  2010     2   2    tmin   144
41  2010    11   2    tmax   313
42  2010    11   2    tmin   163
```

dcast

```
tidy <- dcast(raw, year + month + day ~ element,
  value.var = ”value”)
```

data frame to reshape

```
tidy <- dcast(raw, year + month + day ~ element,
  value.var = "value")
```

```
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
```

data frame to reshape

column(s) to keep as is

```
tidy <- dcast(raw, year + month + day ~ element,
  value.var = "value")
```

```
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
2010     1  30
2010     2   2
2010     2   3
2010     2  11
2010     2  23
2010     3   5
```

data frame to reshape

column(s) to keep as is

~

```
tidy <- dcast(raw, year + month + day ~ element,
  value.var = "value")
```

```
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
2010     1  30
2010     2   2
2010     2   3
2010     2  11
2010     2  23
2010     3   5
```

data frame to reshape

column(s) to keep as is

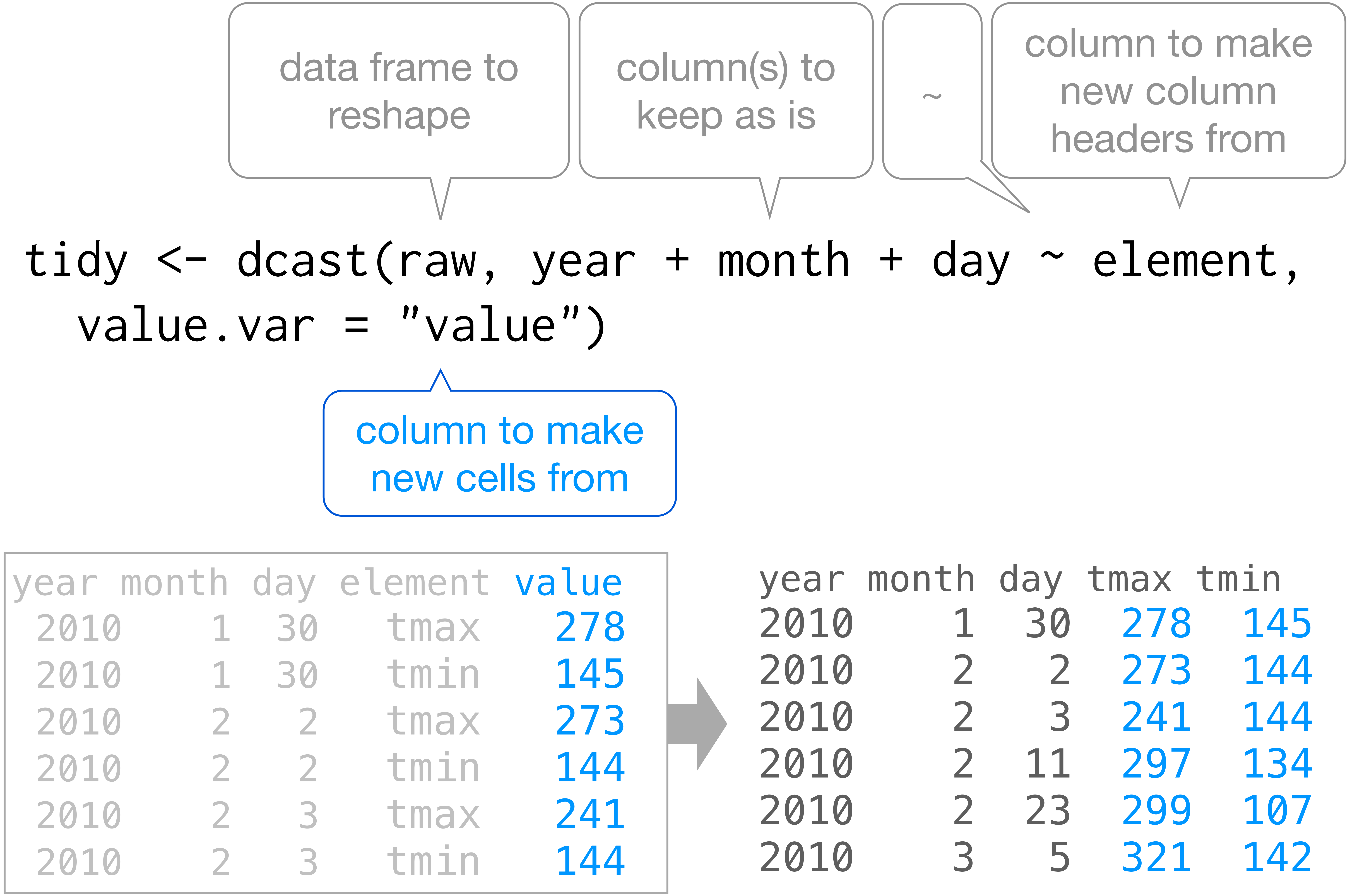
~

column to make new column headers from

```
tidy <- dcast(raw, year + month + day ~ element,
  value.var = "value")
```

```
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    –    –
2010     2   2    –    –
2010     2   3    –    –
2010     2  11    –    –
2010     2  23    –    –
2010     3   5    –    –
```



data frame to reshape

column(s) to keep as is

~

column to make new column headers from

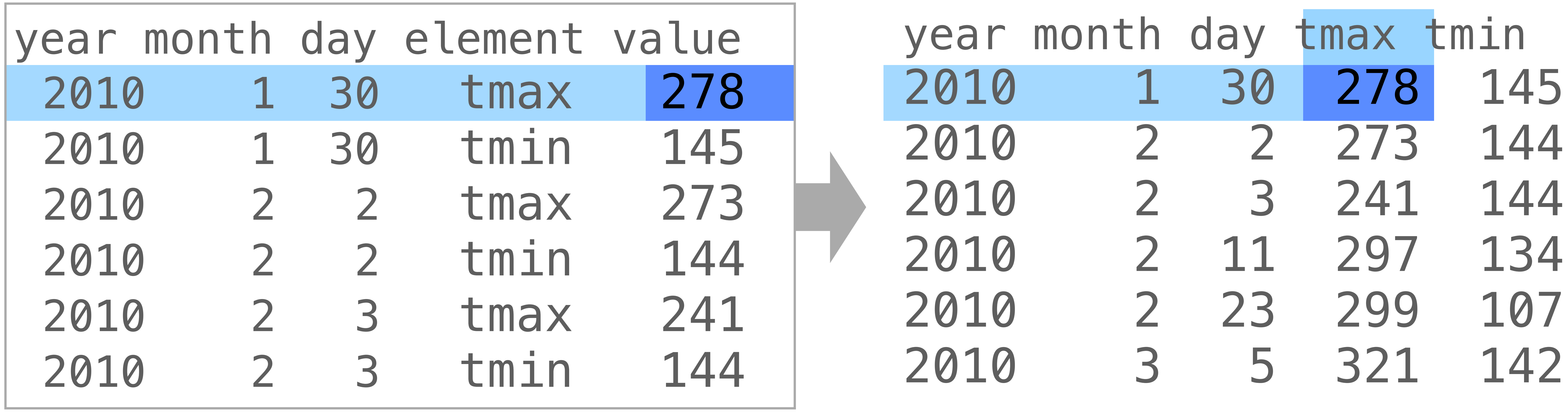
```r
tidy <- dcast(raw, year + month + day ~ element,
  value.var = "value")
```

column to make new cells from

```
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
```

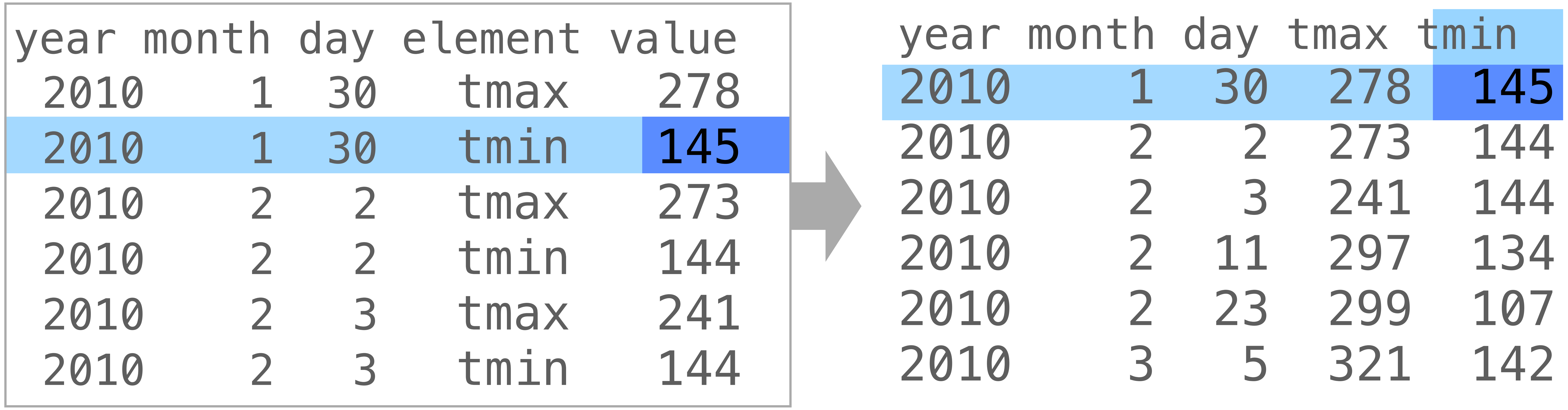
Every combination of values is retained



```
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
```

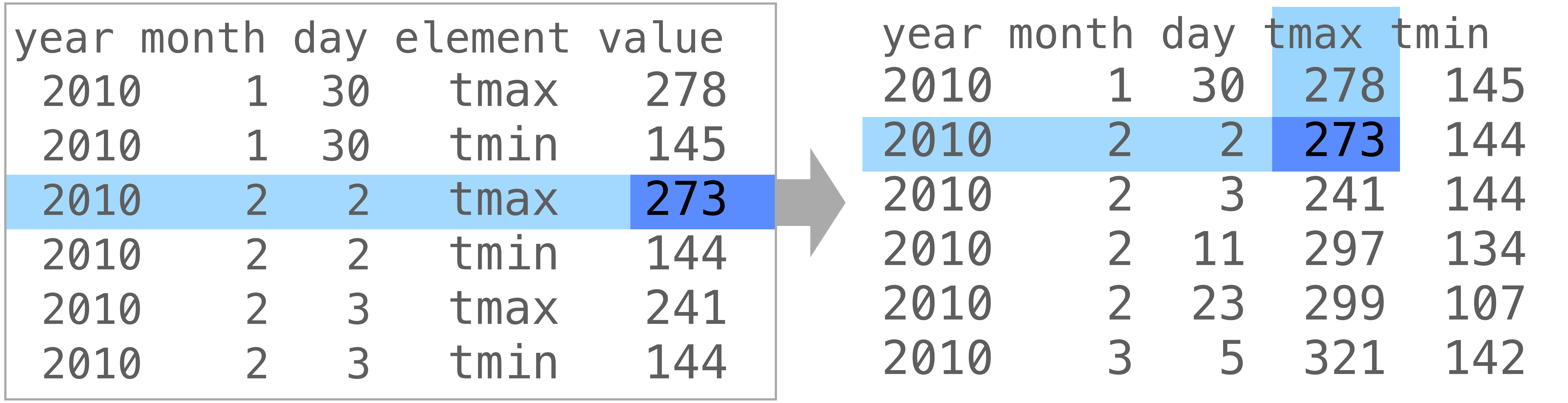
Every combination of values is retained



```
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
```

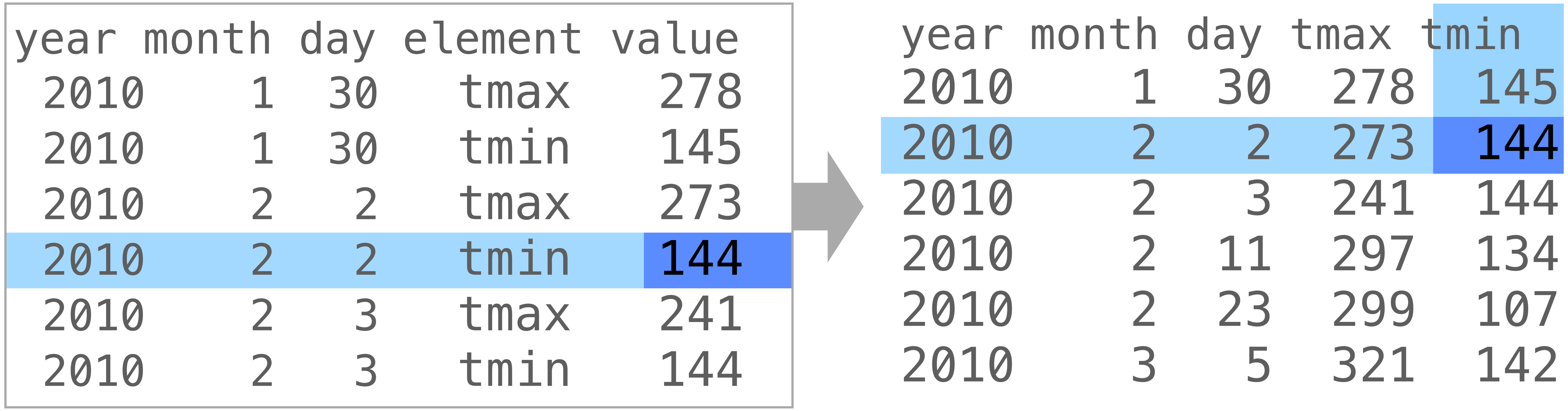
Every combination of values is retained



```
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
```

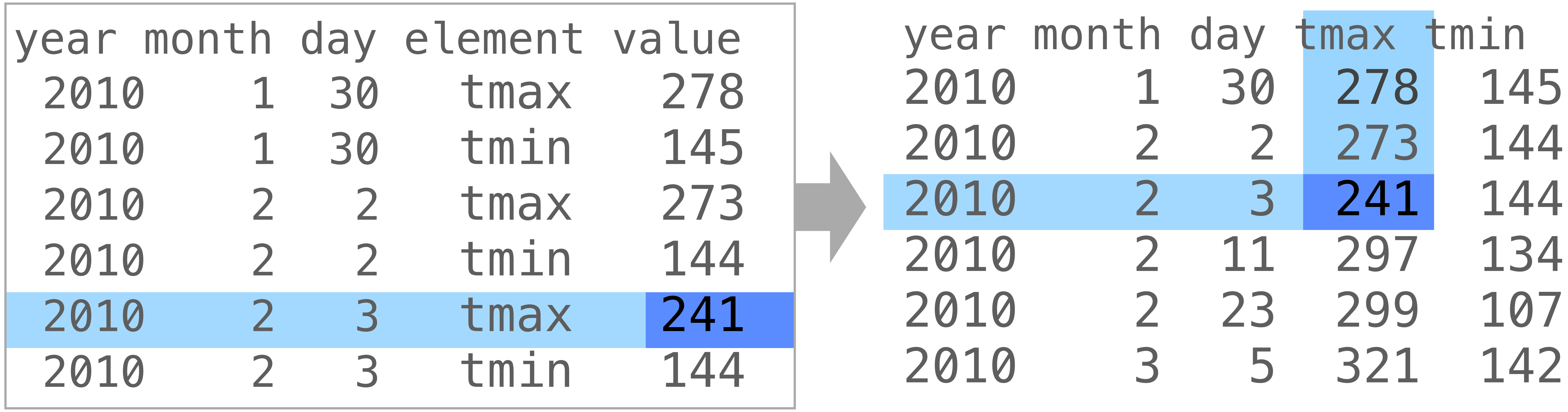
Every combination of values is retained



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

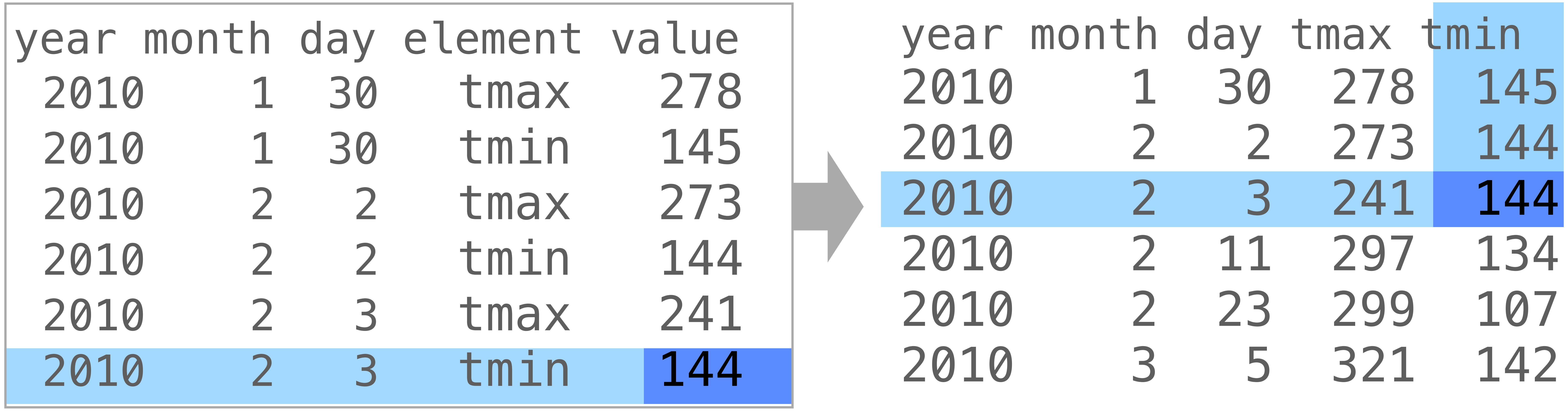
Every combination of values is retained



```
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
```

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(”data/titanic2.csv”,
  stringsAsFactors = FALSE)
```

```
head(titanic2)
#  class   age     fate male female
#    1st adult perished  118      4
#    1st adult survived   57    140
#    1st child perished    0      0
#    1st child survived    5      1
#    2nd adult perished  154     13
#    2nd adult survived   14     80
```

Your turn

$$\text{survival rate} = \frac{\#\ \text{survived}}{\#\ \text{survived} + \#\ \text{perished}}$$

Compute the survival rate of each unique group of age, class, and gender.

You will need to practice everything you've learned so far to tidy titanic2.

(Hint: the equation above requires survived and perished to be separate variables)

05:00

Step 1

```
tidy <- melt(titanic2, id = c("class", "age", "fate"),
  variable.name = "gender")
head(tidy)
# class   age     fate gender value
#   1st adult perished   male   118
#   1st adult survived   male    57
#   1st child perished   male     0
#   1st child survived   male     5
#   2nd adult perished   male   154
#   2nd adult survived   male    14
```

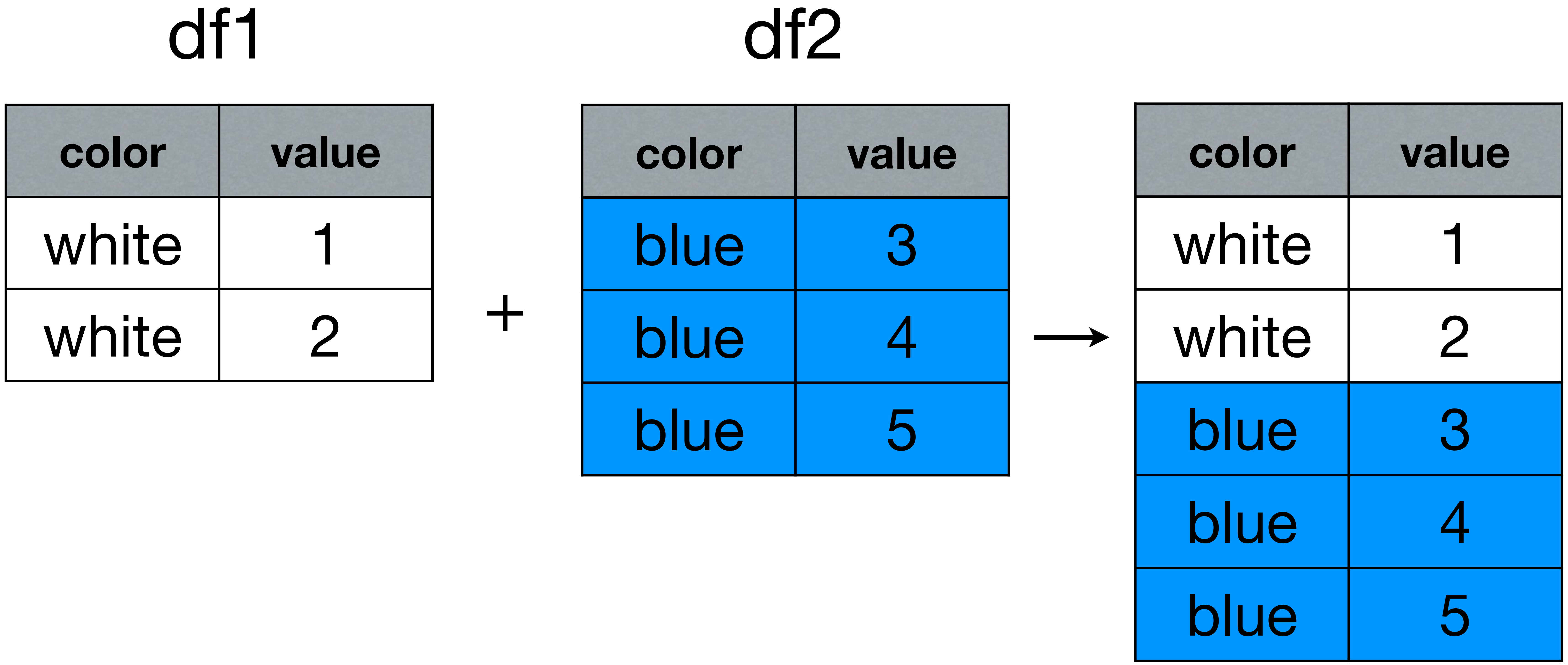
Step 2

```
tidy <- dcast(tidy, class + age + gender ~ fate,
  value.var = ”value”)
head(tidy)
#  class   age gender perished survived
#    1st adult   male      118       57
#    1st adult female        4      140
#    1st child   male        0        5
#    1st child female        0        1
#    2nd adult   male      154       14
#    2nd adult female       13       80
```

Step 3

```
tidy$rate <- round(tidy$survived /
  (tidy$survived + tidy$perished), 2)
head(tidy)
# class   age gender perished survived rate
#   1st adult   male      118       57 0.33
#   1st adult female        4      140 0.97
#   1st child   male        0        5 1.00
#   1st child female        0        1 1.00
#   2nd adult   male      154       14 0.08
#   2nd adult female       13       80 0.86
```

Data split across many files



df1

color	value
white	1
white	2

+

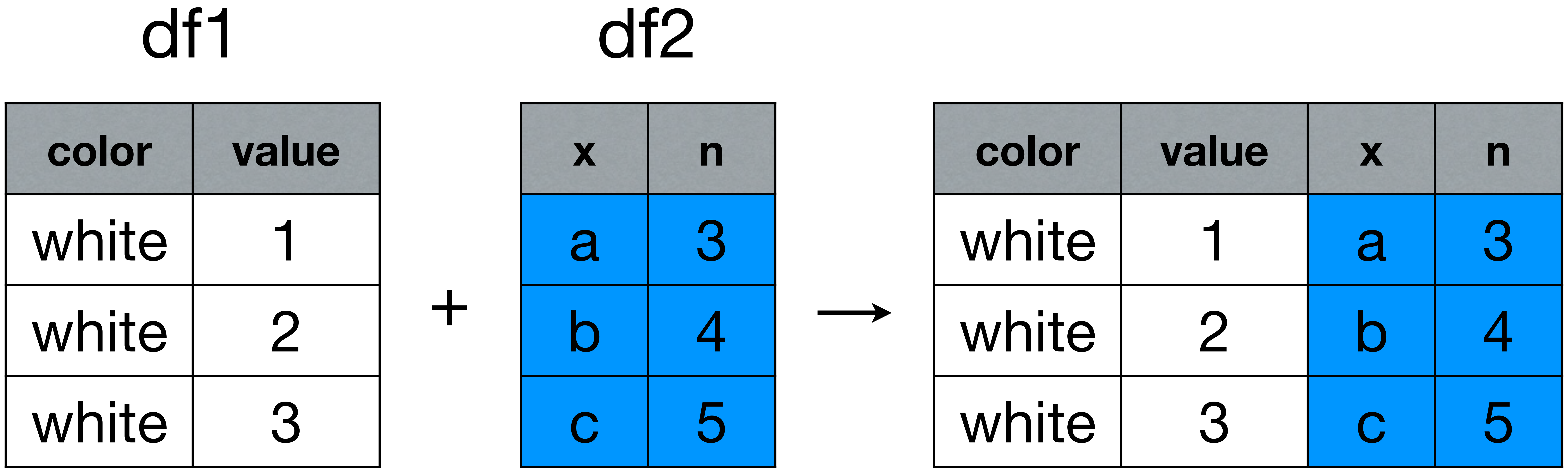
df2

color	value
blue	3
blue	4
blue	5

→

color	value
white	1
white	2
blue	3
blue	4
blue	5

```
rbind(df1, df2)
```



df1

color	value
white	1
white	2
white	3

+

df2

x	n
a	3
b	4
c	5

→

color	value	x	n
white	1	a	3
white	2	b	4
white	3	c	5

```
cbind(df1, df2)
```

Saving data

Saving data

```
# For long-term storage
write.csv(tidy, file = ”tidy.csv”,
  row.names = FALSE)

# For short-term caching
# Preserves factors etc.
saveRDS(tidy, ”tidy.rds”)
tidy2 <- readRDS(”tidy.rds”)
```

Data will be saved in your working directory

.csv	.rds
`read.csv()`	`readRDS()`
`write.csv(row.names = FALSE)`	`saveRDS()`
Only data frames	Any R object
Can be read by any program	Only by R
Long term storage	Short term caching of expensive computations

```
# Easy to store compressed files to save space:
write.csv(tidy, file = bzfile("tidy.csv.bz2”),
  row.names = FALSE)

# Reading is even easier:
tidy3 <- read.csv("tidy.csv.bz2”)

# Files stored with saveRDS() are automatically
# compressed.
```