Tidy Data
(CMP595 PPGC/INF/UFRGS)

Lucas Mello Schnorr, Jean-Marc Vincent

INF/UFRGS

Porto Alegre, Brazil – October 2018

What is tidy data?

Data that makes data analysis easy

- To model
- To visualize
- To manipulate

The tidying process can be seen as a clean-up procedure.

Motivation

	Pregnant	Not pregnant
Male	0	5
Female	1	4

Motivation

	Pregnant	Not pregnant
Male	0	5
Female	1	4

How many variables are in this data set?
What are they?

Tidy data

Each variable is a column.
Each row is an observation.
Each table/file is a data set.

Tidy data

Each variable is a column.
Each row is an observation.
Each table/file is a data set.

Pregnant	Gender	Count
yes	female	1
yes	male	0
no	female	4
no	male	5

Not tidy data

- Column header are values
- Multiple variables are stored in one column
- Variables in rows and columns

Not tidy data

- Column header are values
- Multiple variables are stored in one column
- Variables in rows and columns

- Messy data is common

Column header are values

Income Distribution within U.S. religious groups
by the Pew Forum on Religious and Public life

religion	L30k	30k-50k	50k-100k	M100k	sample
Buddhist	36%	18%	32%	13%	233
Catholic	36%	19%	26%	19%	6137
Evangelical Protestant	35%	22%	28%	14%	7462
Hindu	17%	13%	34%	36%	172
Historically Black Protestant	53%	22%	17%	8%	1704
Jehovah's Witness	48%	25%	22%	4%	208
Jewish	16%	15%	24%	44%	708
Mainline Protestant	29%	20%	28%	23%	5208
Mormon	27%	20%	33%	20%	594
Muslim	34%	17%	29%	20%	205
Orthodox Christian	18%	17%	36%	29%	155
Unaffiliated	33%	20%	26%	21%	6790

Column header are values

Income Distribution within U.S. religious groups
by the Pew Forum on Religious and Public life

religion	L30k	30k-50k	50k-100k	M100k	sample
Buddhist	36%	18%	32%	13%	233
Catholic	36%	19%	26%	19%	6137
Evangelical Protestant	35%	22%	28%	14%	7462
Hindu	17%	13%	34%	36%	172
Historically Black Protestant	53%	22%	17%	8%	1704
Jehovah's Witness	48%	25%	22%	4%	208
Jewish	16%	15%	24%	44%	708
Mainline Protestant	29%	20%	28%	23%	5208
Mormon	27%	20%	33%	20%	594
Muslim	34%	17%	29%	20%	205
Orthodox Christian	18%	17%	36%	29%	155
Unaffiliated	33%	20%	26%	21%	6790

What are the variables in this data set?

Tidying data

Using the gather package (tidyr package)

religion	L30k	30k-50k	50k-100k	M100k	sample
Buddhist	36%	18%	32%	13%	233
Catholic	36%	19%	26%	19%	6137
Evangelical Protestant	35%	22%	28%	14%	7462

```
library(tidyverse);
pew <- as.data.frame(pew)
gather(pew, key, value, -religion, -sample)
```

Tidying data

Using the gather package (tidyr package)

religion	L30k	30k-50k	50k-100k	M100k	sample
Buddhist	36%	18%	32%	13%	233
Catholic	36%	19%	26%	19%	6137
Evangelical Protestant	35%	22%	28%	14%	7462

```
library(tidyverse);
pew <- as.data.frame(pew)
gather(pew, key, value, -religion, -sample)
```

religion	sample	key	value
Buddhist	233	L30k	36%
Catholic	6,137	L30k	36%
Evangelical Protestant	7,462	L30k	35%
Hindu	172	L30k	17%
Historically Black Protestant	1,704	L30k	53%
Jehovah's Witness	208	L30k	48%
Jewish	708	L30k	16%
Mainline Protestant	5,208	L30k	29%
Mormon	594	L30k	27%
Muslim	205	L30k	34%
Orthodox Christian	155	L30k	18%
Unaffiliated	6,790	L30k	33%
...			

Importing tidy data

```
library(readr);
df <- read_delim(file="name_of_the_file.csv",
                 delim="|",
                 col_names=FALSE);
?read_delim
?read_csv
```

References

- Tidy Data, by Hadley Wickham
  - See Section 2, or check directly the Table 3
- The Elements of Data Analytic Style, by Jeff Leek
  - See Section 3.4, Page 12.
- Hadley Wickham presentation `http://vimeo.com/3372755`
  `http://stat405.had.co.nz/lectures/18-tidy-data.pdf`