Why tidy data matters

Gerard Walsh

About me

- Studied engineering (Mechanical & Mechatronic)
- Transitioning into Data Science (@Exegetic)

Data

- 80% of time spent on data preparation
- Repeated processes
- A principled approach?

Tidy data

- Each variable = a column
- Each observation = row
- Each observational unit = table
- By Hadley Wickham

Data

- Relationship between male and female TB cases over time?

	country	year	m04	m514	m014	m1524	m2534	m3544	m4554	m5564
	<fct>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
1	ZW	2003	NA	NA	133	874	3048	2228	981	367
2	ZW	2004	NA	NA	187	833	2908	2298	1056	366
3	ZW	2005	NA	NA	210	837	2264	1855	762	295
4	ZW	2006	NA	NA	215	736	2391	1939	896	348
5	ZW	2007	6	132	138	500	3693	0	716	292
6	ZW	2008	NA	NA	127	614	0	3316	704	263

World Health Organization

M2534 = males aged 25-34

Data

- Each variable != column
- Each row != observation

	country	year	m04	m514	m014	m1524	m2534	m3544	m4554	m5564
	<fct>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>
1	ZW	2003	NA	NA	133	874	3048	2228	981	367
2	ZW	2004	NA	NA	187	833	2908	2298	1056	366
3	ZW	2005	NA	NA	210	837	2264	1855	762	295
4	ZW	2006	NA	NA	215	736	2391	1939	896	348
5	ZW	2007	6	132	138	500	3693	0	716	292
6	ZW	2008	NA	NA	127	614	0	3316	704	263

Data

- “gather” each column into variable

```
tb <- tb %>%
  gather("raw", "count", -country, -year, na.rm = TRUE)
```

Data

- Much “tidier”
- Multiple variables in one column

	country	year	raw	count
	<fct>	<int>	<chr>	<int>
1	AD	2005	m04	0
2	AD	2006	m04	0
3	AD	2008	m04	0
4	AE	2006	m04	0
5	AE	2007	m04	0
6	AE	2008	m04	0
7	AG	2007	m04	0
8	AL	2005	m04	0
9	AL	2006	m04	1
10	AL	2007	m04	0

Data

- Multiple variables in one column

```
tb <- tb %>%
  separate(raw, c("sex","ages"), 1)
```

Data

- Tidy!

	country	year	sex	age	count
	<fct>	<int>	<chr>	<fct>	<int>
1	AD	1996	f	0-14	0
2	AD	1997	f	0-14	0
3	AD	1999	f	0-14	0
4	AD	2002	f	0-14	0
5	AD	2003	f	0-14	0
6	AD	2004	f	0-14	0
7	AD	2005	f	0-14	0
8	AD	2006	f	0-14	0
9	AD	2008	f	0-14	0
10	AD	1996	f	15-24	1

Why?

- Data exploration
- Standardize data analysis tools
  - ! for junior data scientists
- (promise not sponsored by tidyverse)

Exploration

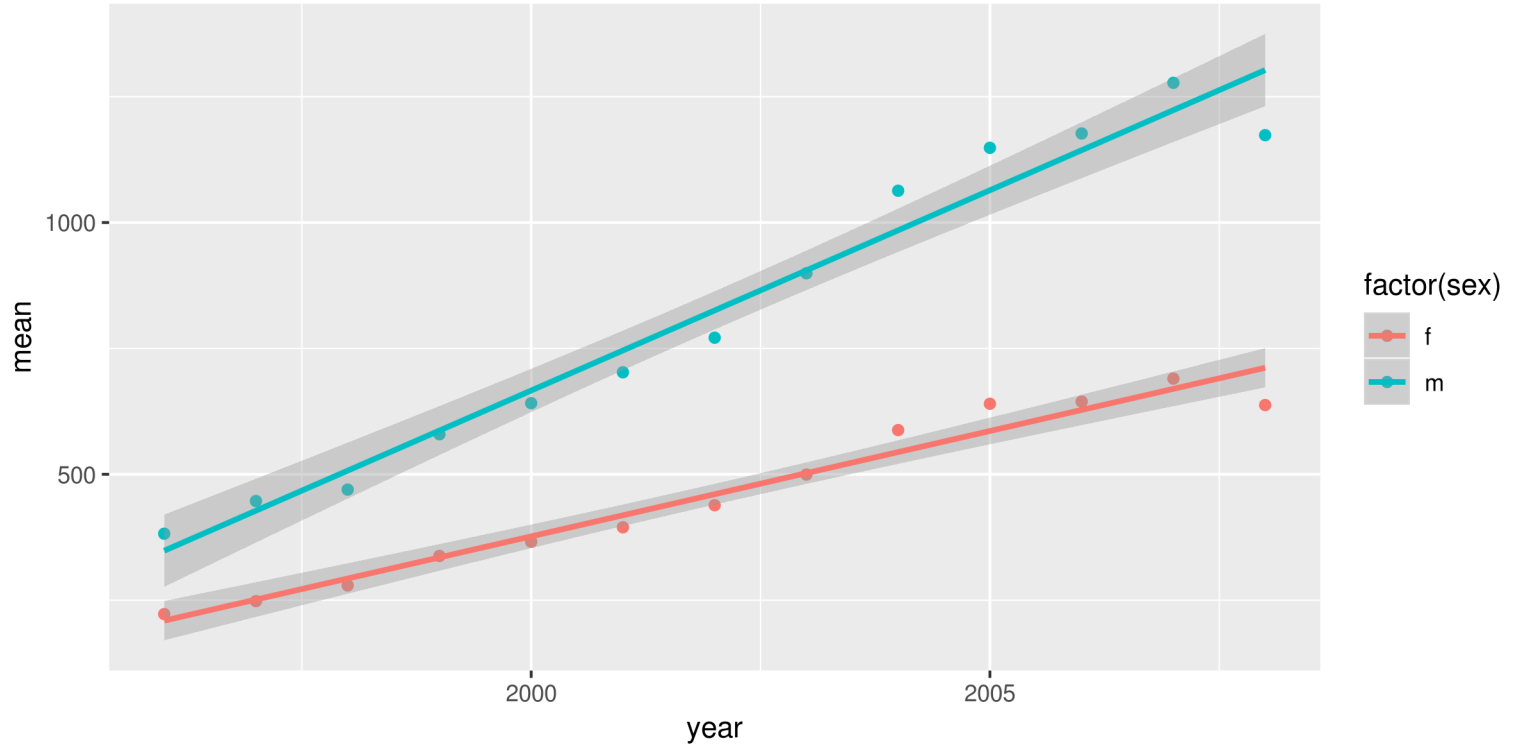
- Relationship between male and female TB cases over time?

```
tb <- tb %>%
  group_by(year, sex) %>%
  summarise(mean = mean(count, na.rm = TRUE))

ggplot(tb, aes(x = year, y = mean, color = factor(sex))) +
  geom_point() +
  stat_smooth(method = "lm")
```

Exploration

- Relationship between male and female TB cases over time?



Take home

- Messy data makes for messy code
- Learn a few packages, and learn them well!

Contact

@gerardlwalsh

@gerrywalsh