

Tidy up your data: **pivot tables**

Dony Indarto

School of Biological, Earth and Environmental Sciences (BEES), UNSW

Why 'tidy' up?

- Consistent structure -> simplify thought process
- Default format expected by stats software



Tidy data rules

- Rows are **observations**
- Columns are **variables**
- Each value has a **cell**

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20095360
Brazil	1999	31737	17206362
Brazil	2000	80488	174504898
China	1999	211258	1272015272
China	2000	216766	128062583

variables

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20095360
Brazil	1999	31737	17206362
Brazil	2000	80488	174504898
China	1999	211258	1272015272
China	2000	216766	128062583

observations

country	year	cases	population
Afghanistan	99	745	19987071
Afghanistan	00	2666	20095360
Brazil	99	31737	17206362
Brazil	00	80488	174504898
China	99	211258	1272015272
China	00	216766	128062583

values

Is this data 'tidy'?

		Columns (variables)			
Rows (observations)		genre	mpaa_rating	domestic_gross	worldwide_gross
{	Evan Almighty	Comedy	PG	100289690	174131329
	Waterworld	Action	PG-13	88246220	264246220
	King Arthur: Legend of the Sword	Adventure	PG-13	39175066	139950708
	47 Ronin	Action	PG-13	38362475	151716815
	Jurassic World: Fallen Kingdom	Action	PG-13	416769345	1304866322
		Cells (values)			



Which one is 'tidy'? #1

Columns are values, not variable names.

religion	<\$10k	\$10-20k	\$20-30k
Buddhist	27	21	30
Jewish	19	19	25
Hindu	1	9	7




religion	income	value
Buddhist	<\$10k	27
Buddhist	\$10-20k	21
Buddhist	\$20-30k	30
Catholic	<\$10k	418
Catholic	\$10-20k	617
Catholic	\$20-30k	732
Hindu	<\$10k	1
Hindu	\$10-20k	9
Hindu	\$20-30k	7
Jewish	<\$10k	19
Jewish	\$10-20k	19
Jewish	\$20-30k	25



Which one is 'tidy'? #2

A.)


Person	treatment	result
John Smith	a	-
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1



B.)

Columns are values, not variable names.


person	treatmenta	treatmentb
John Smith	-	2
Jane Doe	16	11
Mary Johnson	3	1



C.)

Columns are values, not variable names.

person	John Smith	Jane Doe	Mary Johnson
a	-	16	3
b	2	11	1



Which one is 'tidy'? #3



a.) The Fellowship of The Ring

Race	Female	Male
Elf	1229	971
Hobbit	14	3644
Man	0	1995

The Two Towers

Race	Female	Male
Elf	331	513
Hobbit	0	2463
Man	401	3589

Data are stored in multiple tables



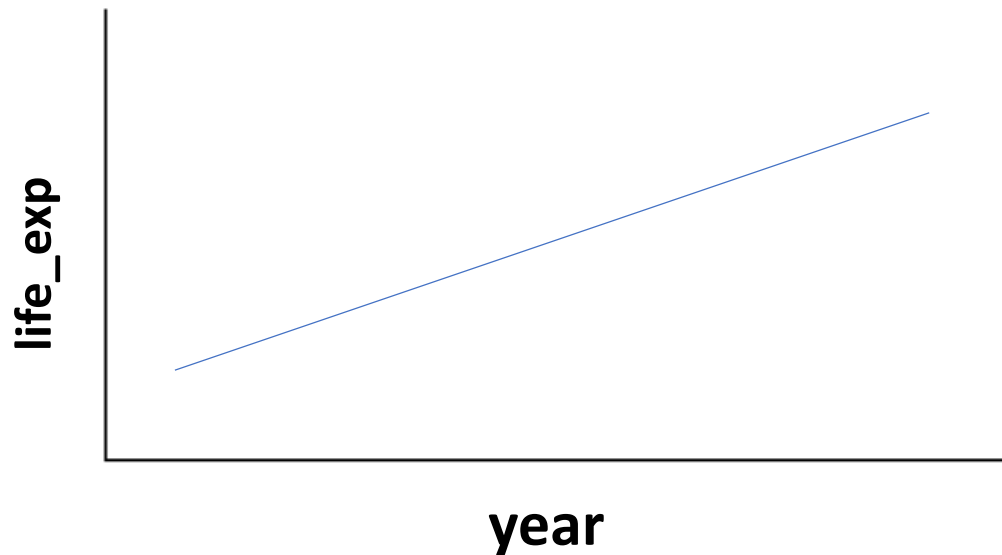
b.)

Film	Gender	Race	Words
The Fellowship Of The Ring	Female	Elf	1229
The Fellowship Of The Ring	Male	Elf	971
The Fellowship Of The Ring	Female	Hobbit	14
The Fellowship Of The Ring	Male	Hobbit	3644
The Fellowship Of The Ring	Female	Man	0
The Fellowship Of The Ring	Male	Man	1995
The Two Towers	Female	Elf	331
The Two Towers	Male	Elf	513
The Two Towers	Female	Hobbit	0
The Two Towers	Male	Hobbit	2463
The Two Towers	Female	Man	401
The Two Towers	Male	Man	3589



Why untidy data is difficult to work with?

Does life expectancy increase over time?



Life expectancy

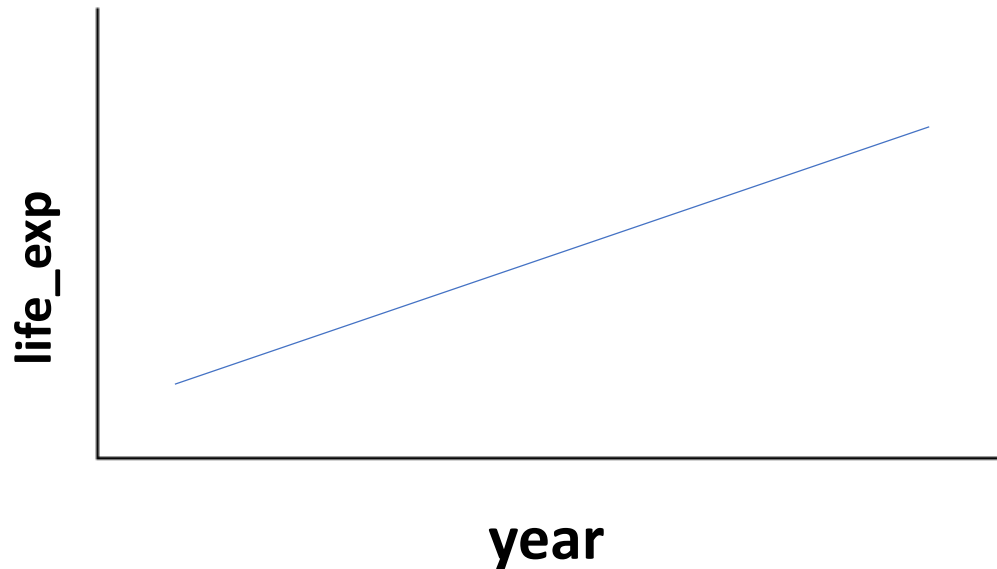
country	1800	1801	1802	...
Afghanistan	28.2	28.2	28.2	...
Albania	35.4	35.4	35.4	...
Algeria	28.8	28.8	28.8	...
Andorra	-	-	-	...
Angola	27	27	27	...
...



```
ggplot(data, aes(x = year, y = life_exp) +  
  geom_line())
```



Does life expectancy increase over time?



Life expectancy

country	year	life_exp
Afghanistan	1800	28.2
Afghanistan	1801	28.2
Afghanistan	1802	28.2
Afghanistan	1803	28.2
Afghanistan	1803	28.2
...

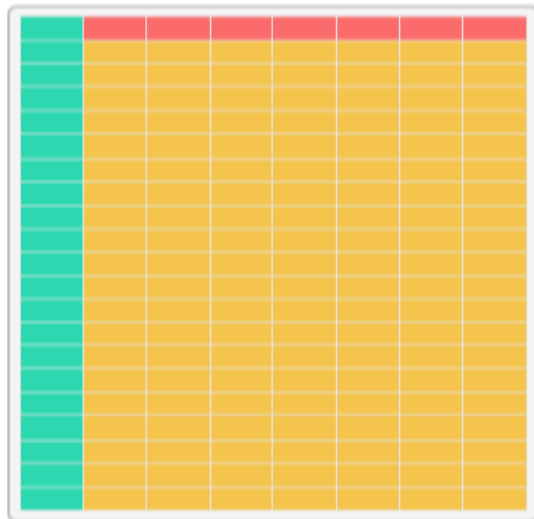


```
ggplot(data, aes(x = year, y = life_exp)) +  
  geom_line()
```



country	1800	1801	1802	...
Afghanistan	28.2	28.2	28.2	...
Albania	35.4	35.4	35.4	...
Algeria	28.8	28.8	28.8	...
Andorra	-	-	-	...
Angola	27	27	27	...
...

country	year	life_exp
Afghanistan	1800	28.2
Afghanistan	1801	28.2
Afghanistan	1802	28.2
Afghanistan	1803	28.2
Afghanistan	1803	28.2
...



Wide format

`pivot_longer()`



`pivot_wider()`



Long format

Make datasets longer (more rows, less columns)

```
pivot_longer(data, cols, names_to, values_to)
```

- `cols`
columns to pivot into longer format
- `names_to`
the name of the column made from the data stored in the col.names
- `values_to`
the name of the column made from the data stored in the cells



Working example: `pivot_longer`

```
library(tidyverse)
le <- read_csv("data/gapminder/life_expectancy_years.csv")
le_long <- pivot_longer(data = le,
                        cols = -country,
                        names_to = "year",
                        values_to = "life_exp")

le_long
```

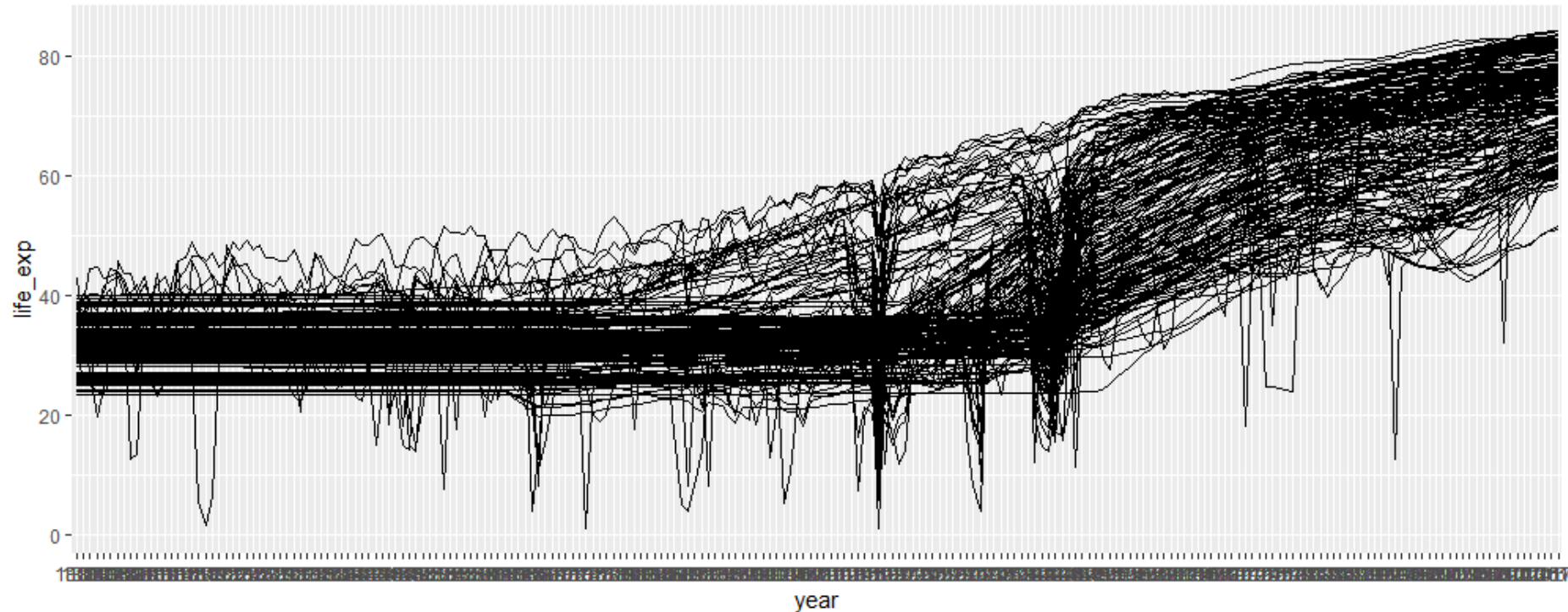
country	year	life_exp
Afghanistan	1800	28.2
Afghanistan	1801	28.2
Afghanistan	1802	28.2
Afghanistan	1803	28.2
Afghanistan	1803	28.2
...



Does life expectancy increase over time?

```
ggplot(le_long,  
  aes(x = year, y = life_exp, group = country)) +  
  geom_line() +  
  theme(legend.position = "none")
```

#data
#aesthetics
#type of plot
#remove legend



Make datasets wider (more columns, less rows)

```
pivot_wider(data, names_from, values_from)
```

- `names_from`
which column to get the column names from
- `values_from`
which column to get the values from



Working example: `pivot_wider`

```
le_wide <- pivot_wider(data = le_long,  
                       names_from = "year",  
                       values_from = "life_exp")  
  
le_wide
```

country	1800	1801	1802	...
Afghanistan	28.2	28.2	28.2	...
Albania	35.4	35.4	35.4	...
Algeria	28.8	28.8	28.8	...
Andorra	-	-	-	...
Angola	27	27	27	...
...

Exercise

Make these data tidy:

- `children_per_woman_total_fertility.csv`
- `income_per_person_gdppercapita_ppp_inflation_adjusted.csv`



References

- Goldsmith, J. (2017). *Tidy data*. [online]. Available at: https://p8105.com/tidy_data.html
- Wickham, H., 2014. Tidy data. *Journal of Statistical Software*, 59(10), pp.1-23.
- Wickham, H. and Henry, L. (2020). *Function reference*. [online] Tidyverse.org. Available at: <https://tidyr.tidyverse.org/reference/>