Data Organization in Spreadsheets and Tidy Data

Sahir Bhatnagar

Department of Diagnostic Radiology
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

sahir.bhatnagar@mcgill.ca
https://sahirbhatnagar.com/

February 27, 2019

Outline

- Collect –> Data in Spreadsheets
- Format –> Tidy Data
- Clean –> Data manipulation

How to organize your data

Reference

Fig.: Data Organization in Spreadsheets

Golden Rules

1. **Be consistent**: “male”, “Male”, “MALE”, “ male”
2. **Dates in this format only**: YYYY-MM-DD
3. **Do not leave cells empty**: use **NA**
4. **Put just one thing in a cell**: “45 grams” $\rightarrow$ “45”
5. **Subjects as rows and variables as columns**
6. **Create a data dictionary**
7. **Do not include calculations in the raw data files**
8. **Do not use font color or highlighting as data**
9. **Choose good names for variables**
10. **Make backups**
11. **Save the data in plain text files**: .txt, .csv

Be consistent

- The first rule of data organization is be consistent. Whatever you do, do it consistently.
- Use consistent codes for categorical variables. Avoid “male”, “Male”, “MALE”

Dates

- Always use the YYYY-MM-DD format

Do not leave cells empty

- An empty cell should always be filled with NA.
- Use a consistent fixed code for any missing values. Do not use 999, -999, N/A
- Be careful about extra spaces within cells.
- A blank cell is different than a cell that contains a single space.
- “male” is different from “ male ” (i.e., with spaces at the beginning and end).

Subjects as rows and variables as columns

- Do not use more than 1 row for the variable names

	A	B	C	D	E
1	id	sex	glucose	insulin	triglyc
2	101	Male	134.1	0.60	273.4
3	102	Female	120.0	1.18	243.6
4	103	Male	124.8	1.23	297.6
5	104	Male	83.1	1.16	142.4
6	105	Male	105.2	0.73	215.7

Fig.: Example of a properly formatted dataset

Create a data dictionary

- **name** exact variable name as in the data file, **plot_name** is the name used for plot labels, **description** is longer explanation of what the variable means

	A	B	C	D
1	name	plot_name	group	description
2	mouse	Mouse	demographic	Animal identifier
3	sex	Sex	demographic	Male (M) or Female (F)
4	sac_date	Date of sac	demographic	Date mouse was sacrificed
5	partial_inflation	Partial inflation	clinical	Indicates if mouse showed partial pancreatic inflation
6	coat_color	Coat color	demographic	Coat color, by visual inspection
7	crumblers	Crumblers	clinical	Indicates if mouse stored food in their bedding
8	diet_days	Days on diet	clinical	Number of days on high–fat diet

Fig.: Data dictionary

Put just one thing in a cell

- Avoid “45 grams”. 45 should be the value of the cell, and grams should be in the data dictionary

Choose good names for variables

Max_temp_C	MaxTemp	Maximum Temp (°C)
Precipitation_mm	Precipitation	precmm
Mean_year_growth	MeanYearGrowth	Mean growth/year
sex	sex	M/F
weight	weight	w.
cell_type	CellType	Cell type
Observation_01	first_observation	1st Obs.

Fig.: Comparison of variable names

Do not use font color or highlighting as data

	A	B	C
1	id	date	glucose
2	101	2015–06–14	149.3
3	102	2015–06–14	95.3
4	103	2015–06–18	97.5
5	104	2015–06–18	1.1
6	105	2015–06–18	108.0
7	106	2015–06–20	149.0
8	107	2015–06–20	169.4

Fig.: Not a good idea

Do not use font color or highlighting as data

	A	B	C
1	id	date	glucose
2	101	2015–06–14	149.3
3	102	2015–06–14	95.3
4	103	2015–06–18	97.5
5	104	2015–06–18	1.1
6	105	2015–06–18	108.0
7	106	2015–06–20	149.0
8	107	2015–06–20	169.4

	A	B	C	D
1	id	date	glucose	outlier
2	101	2015–06–14	149.3	FALSE
3	102	2015–06–14	95.3	FALSE
4	103	2015–06–18	97.5	FALSE
5	104	2015–06–18	1.1	TRUE
6	105	2015–06–18	108.0	FALSE
7	106	2015–06–20	149.0	FALSE
8	107	2015–06–20	169.4	FALSE

Fig.: Think of the color as another variable

Save the data in plain text files

	A	B	C	D	E
1	id	sex	glucose	insulin	triglyc
2	101	Male	134.1	0.60	273.4
3	102	Female	120.0	1.18	243.6
4	103	Male	124.8	1.23	297.6
5	104	Male	83.1	1.16	142.4
6	105	Male	105.2	0.73	215.7

```
id,sex,glucose,insulin,triglyc
101,Male,134.1,0.60,273.4
102,Female,120.0,1.18,243.6
103,Male,124.8,1.23,297.6
104,Male,83.1,1.16,142.4
105,Male,105.2,0.73,215.7
```

Fig.: Save as .csv file

Example 1

	A	B	C
1	Date	Assay date	Weight
2		12/9/05	54.9
3		12/9/05	45.3
4	12/6/2005	e	47
5		e	45.7
6		e	52.9
7		1/11/2006	46.1
8		1/11/2006	38.6

Fig.: Is this good or bad. Why?

Example 1

	A	B	C
1	Date	Assay date	Weight
2		12/9/05	54.9
3		12/9/05	45.3
4	12/6/2005	e	47
5		e	45.7
6		e	52.9
7		1/11/2006	46.1
8		1/11/2006	38.6

Fig.: Is this good or bad. Why?

Verdict: Bad

Example 2

	A	B	C
1	id	date	glucose
2	101	2015–06–14	149.3
3	102		95.3
4	103	2015–06–18	97.5
5	104		117.0
6	105		108.0
7	106	2015–06–20	149.0
8	107		169.4

Fig.: Is this good or bad. Why?

Example 2

	A	B	C
1	id	date	glucose
2	101	2015–06–14	149.3
3	102		95.3
4	103	2015–06–18	97.5
5	104		117.0
6	105		108.0
7	106	2015–06–20	149.0
8	107		169.4

Fig.: Is this good or bad. Why?

Verdict: Bad

Example 3

	A	B	C	D	E	F	G	H	I
1		1 min				5 min			
2	strain	normal		mutant		normal		mutant	
3	A	147	139	166	179	334	354	451	474
4	B	246	240	178	172	514	611	412	447

Fig.: Is this good or bad. Why?

Example 3

	A	B	C	D	E	F	G	H	I
1		1 min				5 min			
2	strain	normal		mutant		normal		mutant	
3	A	147	139	166	179	334	354	451	474
4	B	246	240	178	172	514	611	412	447

Fig.: Is this good or bad. Why?

Verdict: Bad

Example 3 (continued)

	A	B	C	D	E
1	strain	genotype	min	replicate	response
2	A	normal	1	1	147
3	A	normal	1	2	139
4	B	normal	1	1	246
5	B	normal	1	2	240
6	A	mutant	1	1	166
7	A	mutant	1	2	179
8	B	mutant	1	1	178
9	B	mutant	1	2	172
10	A	normal	5	1	334
11	A	normal	5	2	354
12	B	normal	5	1	514
13	B	normal	5	2	611
14	A	mutant	5	1	451
15	A	mutant	5	2	474
16	B	mutant	5	1	412
17	B	mutant	5	2	447

Fig.: Is this good or bad. Why?

Example 3 (continued)

	A	B	C	D	E
1	strain	genotype	min	replicate	response
2	A	normal	1	1	147
3	A	normal	1	2	139
4	B	normal	1	1	246
5	B	normal	1	2	240
6	A	mutant	1	1	166
7	A	mutant	1	2	179
8	B	mutant	1	1	178
9	B	mutant	1	2	172
10	A	normal	5	1	334
11	A	normal	5	2	354
12	B	normal	5	1	514
13	B	normal	5	2	611
14	A	mutant	5	1	451
15	A	mutant	5	2	474
16	B	mutant	5	1	412
17	B	mutant	5	2	447

Fig.: Is this good or bad. Why?

Verdict: Good

Example 4

	A	B	C	D	E	F	G	H	I	J	K
1			week 4			week 6			week 8		
2	Mouse ID	SEX	date	weight	glucose	date	weight	glucose	date	weight	glucose
3	3005	M	3/30/2007	19.3	635	4/11/2007	31	460.7	4/27/2007	39.6	530.2
4	3017	M	10/6/2006	25.9	202.4	10/19/2006	45.1	384.7	11/3/2006	57.2	458.7
5	3434	F	11/22/2006	26.6	238.9	12/6/2006	45.9	378	12/22/2006	56.2	409.8
6	3449	M	1/5/2007	27.5	121	1/19/2007	42.9	191.3	2/2/2007	56.7	182.5
7	3499	F	1/5/2007	19.8	220.2	1/19/2007	36.6	556.9	2/2/2007	43.6	446

Fig.: Is this good or bad. Why?

Example 4

	A	B	C	D	E	F	G	H	I	J	K
1			week 4			week 6			week 8		
2	Mouse ID	SEX	date	weight	glucose	date	weight	glucose	date	weight	glucose
3	3005	M	3/30/2007	19.3	635	4/11/2007	31	460.7	4/27/2007	39.6	530.2
4	3017	M	10/6/2006	25.9	202.4	10/19/2006	45.1	384.7	11/3/2006	57.2	458.7
5	3434	F	11/22/2006	26.6	238.9	12/6/2006	45.9	378	12/22/2006	56.2	409.8
6	3449	M	1/5/2007	27.5	121	1/19/2007	42.9	191.3	2/2/2007	56.7	182.5
7	3499	F	1/5/2007	19.8	220.2	1/19/2007	36.6	556.9	2/2/2007	43.6	446

Fig.: Is this good or bad. Why?

Verdict: Bad

Example 4 (continued)

	A	B	C	D	E	F
1	mouse_id	sex	week	date	glucose	weight
2	3005	M	4	3/30/2007	19.3	635
3	3005	M	6	4/11/2007	31	460.7
4	3005	M	8	4/27/2007	39.6	530.2
5	3017	M	4	10/6/2006	25.9	202.4
6	3017	M	6	10/19/2006	45.1	384.7
7	3017	M	8	11/3/2006	57.2	458.7
8	3434	F	4	11/22/2006	26.6	238.9
9	3434	F	6	12/6/2006	45.9	378
10	3434	F	8	12/22/2006	56.2	409.8
11	3449	M	4	1/5/2007	27.5	121
12	3449	M	6	1/19/2007	42.9	191.3
13	3449	M	8	2/2/2007	56.7	182.5
14	3499	F	4	1/5/2007	19.8	220.2
15	3499	F	6	1/19/2007	36.6	556.9
16	3499	F	8	2/2/2007	43.6	446

Fig.: Is this good or bad. Why?

Example 4 (continued)

	A	B	C	D	E	F
1	mouse_id	sex	week	date	glucose	weight
2	3005	M	4	3/30/2007	19.3	635
3	3005	M	6	4/11/2007	31	460.7
4	3005	M	8	4/27/2007	39.6	530.2
5	3017	M	4	10/6/2006	25.9	202.4
6	3017	M	6	10/19/2006	45.1	384.7
7	3017	M	8	11/3/2006	57.2	458.7
8	3434	F	4	11/22/2006	26.6	238.9
9	3434	F	6	12/6/2006	45.9	378
10	3434	F	8	12/22/2006	56.2	409.8
11	3449	M	4	1/5/2007	27.5	121
12	3449	M	6	1/19/2007	42.9	191.3
13	3449	M	8	2/2/2007	56.7	182.5
14	3499	F	4	1/5/2007	19.8	220.2
15	3499	F	6	1/19/2007	36.6	556.9
16	3499	F	8	2/2/2007	43.6	446

Fig.: Is this good or bad. Why?

Verdict: Good

Example 5

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	#	M/F (1: male)	DOB (D/M/Y)	Age	Chemonaive (1: yes)		Lesion #	PATH (1: DESMO)		Meta/sync	INDEX DATE	SURVIVAL	DATE OF FOLLOWUP	RECURRENCE	DATE OF RECURRENCE	
2	1	1	9/24/1961	49	1		#001	1		0	6/27/2011	1	6/17/2015	1	1/7/2014	
3	1	1	9/24/1961	49	1		#002	1		0	6/27/2011	1	6/17/2015	1	1/7/2014	
4	1	1	9/24/1961	49	1		#003	0		0	6/27/2011	1	6/17/2015	1	1/7/2014	

Fig.: Is this good or bad. Why?

Example 5

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	#	M/F (1: male)	DOB (D/M/Y)	Age	Chemonaive (1: yes)		Lesion #	PATH (1: DESMO)		Meta/sync	INDEX DATE	SURVIVAL	DATE OF FOLLOWUP	RECURRENCE	DATE OF RECURRENCE	
2	1	1	9/24/1961	49	1		#001	1		0	6/27/2011	1	6/17/2015	1	1/7/2014	
3	1	1	9/24/1961	49	1		#002	1		0	6/27/2011	1	6/17/2015	1	1/7/2014	
4	1	1	9/24/1961	49	1		#003	0		0	6/27/2011	1	6/17/2015	1	1/7/2014	

Fig.: Is this good or bad. Why?

Verdict: Bad

Tidy Data

Tidy data

Journal of Statistical Software

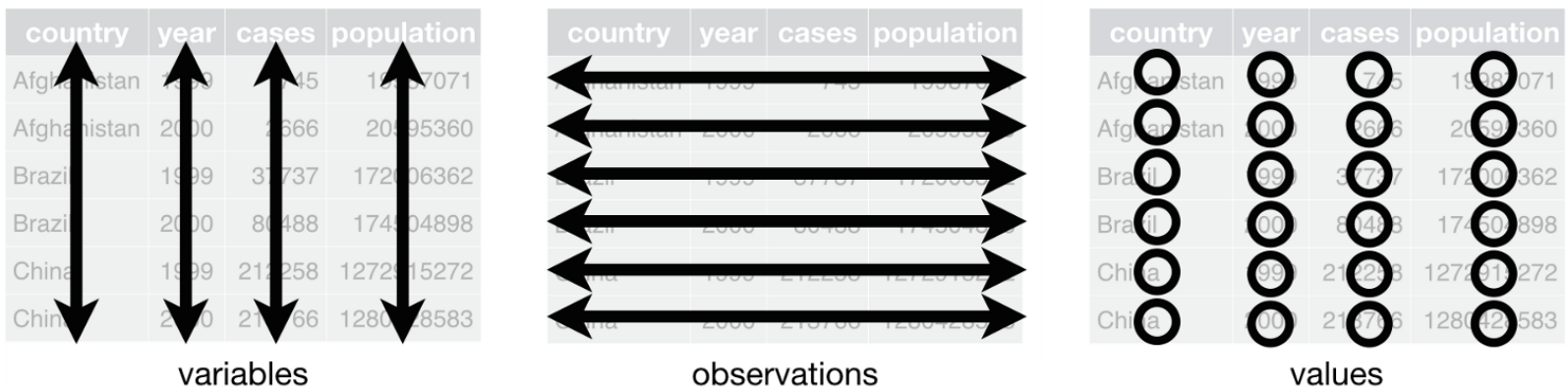
August 2014, Volume 59, Issue 10. *http://www.jstatsoft.org/*

Tidy Data

Hadley Wickham
RStudio

Tidy data

- Each variable forms a column.
- Each observation forms a row.
- Each type of observational units forms a table
- Tidy data is ready for regression routines and plotting



Example: Does a full moon affect behaviour?

- Many people believe that the moon influences the actions of some individuals.
- A study of dementia patients in nursing homes recorded various types of disruptive behaviors every day for 12 weeks.
- Days were classified as moon days if they were in a 3-day period centered at the day of the full moon.
- For each patient, the average number of disruptive behaviors was computed for moon days and for all otherdays.

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26
6	3.67	0.11
7	4.67	0.30

Is it tidy?

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32

Is it tidy?

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32

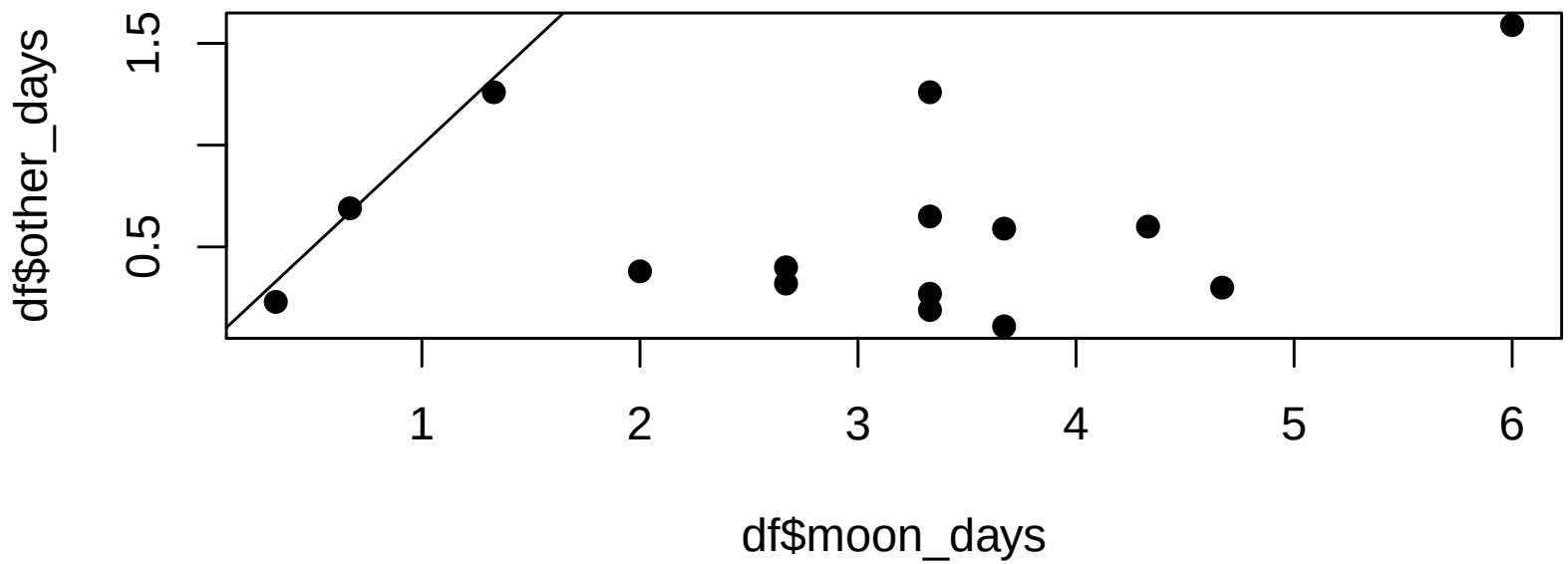
Question: Can I plot the data?

Is it tidy?

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32

Question: Can I plot the data?

```
plot(df$moon_days, df$other_days, pch = 19)
abline(a=0,b=1)
```



Is it tidy?

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26

Is it tidy?

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26

Question: Can I fit a meaningful regression model directly to the variables in the data?

Is it tidy?

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26

Question: Can I fit a meaningful regression model directly to the variables in the data?

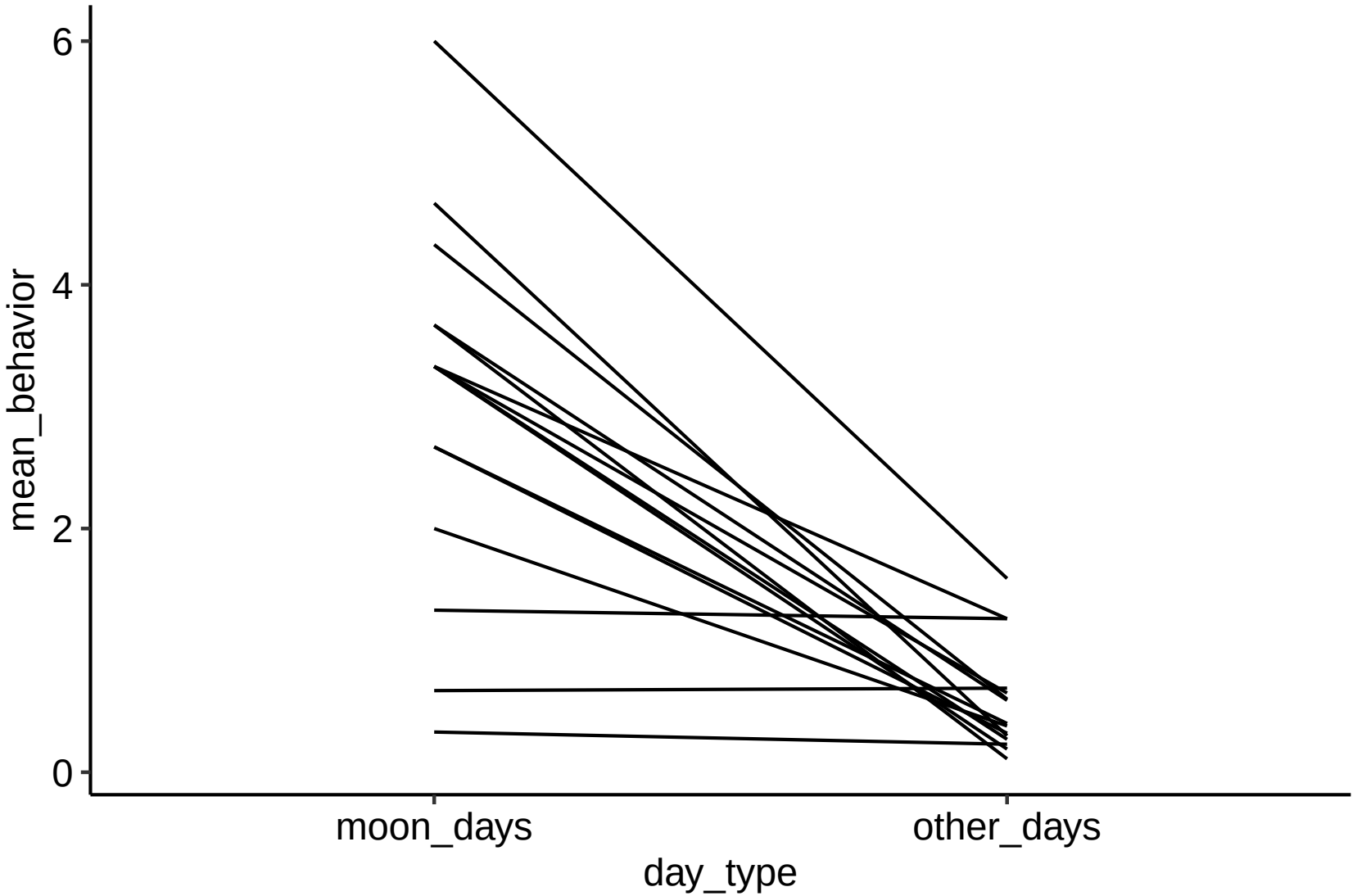
```
Call: lm(formula = moon_days ~ other_days, data = df)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.56       0.66     3.9    0.002
other_days      0.79       0.91     0.9    0.402

Residual standard error: 1.5 on 13 degrees of freedom
Multiple R-squared: 0.055,^^IAdjusted R-squared: -0.018
F-statistic: 0.75 on 1 and 13 DF,  p-value: 0.4
```

Is it tidy?

patient	day_type	mean_behavior
1	moon_days	3.33
1	other_days	0.27
2	moon_days	3.67
2	other_days	0.59
3	moon_days	2.67
3	other_days	0.32
4	moon_days	3.33
4	other_days	0.19
5	moon_days	3.33
5	other_days	1.26

Plotting with tidy data

```
ggformula::gf_line(mean_behavior ~ day_type, group = ~ patient, data = df_t)
```



Regression with tidy data

```
fit <- lme4::lmer(mean_behavior ~ day_type + (1|patient), data = df_t)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: mean_behavior ~ day_type + (1 | patient)
   Data: df_t

REML criterion at convergence: 90.3

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.27236 -0.30142 -0.04023  0.48540  2.44753

Random effects:
 Groups   Name        Variance Std.Dev.
 patient  (Intercept) 0.1563   0.3954
 Residual             1.0659   1.0324
Number of obs: 30, groups:  patient, 15

Fixed effects:
                  Estimate Std. Error t value
(Intercept)         3.0220     0.2854  10.587
day_typeother_days -2.4327     0.3770  -6.453

Correlation of Fixed Effects:
            (Intr)
dy_typthr_d -0.660
```

Not tidy vs. tidy data

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26

patient	day_type	mean_behavior
1	moon_days	3.33
1	other_days	0.27
2	moon_days	3.67
2	other_days	0.59
3	moon_days	2.67
3	other_days	0.32
4	moon_days	3.33
4	other_days	0.19
5	moon_days	3.33
5	other_days	1.26

Not tidy vs. tidy data

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26

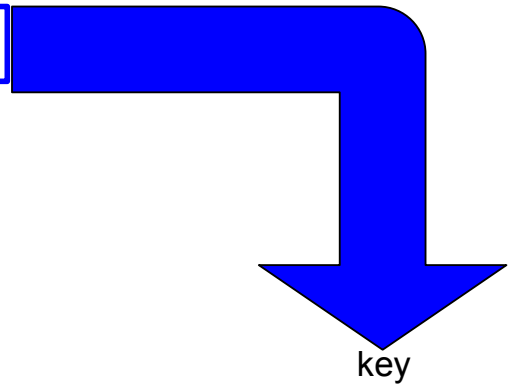
Not tidy

tidy

patient	day_type	mean_behavior
1	moon_days	3.33
1	other_days	0.27
2	moon_days	3.67
2	other_days	0.59
3	moon_days	2.67
3	other_days	0.32
4	moon_days	3.33
4	other_days	0.19
5	moon_days	3.33
5	other_days	1.26

tidyr::gather()

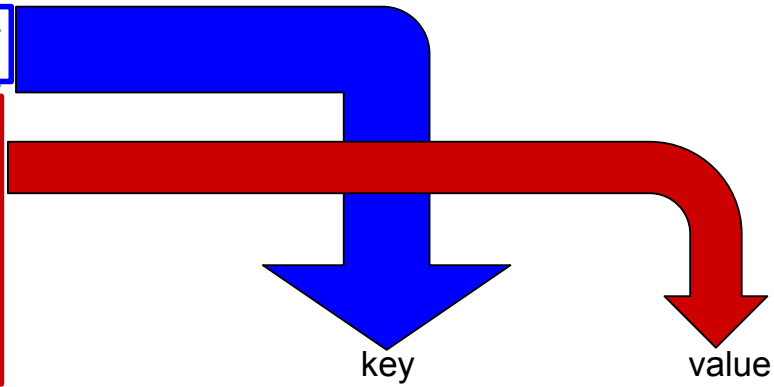
patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26



patient	day_type	mean_behavior
1	moon_days	3.33
1	other_days	0.27
2	moon_days	3.67
2	other_days	0.59
3	moon_days	2.67
3	other_days	0.32
4	moon_days	3.33
4	other_days	0.19
5	moon_days	3.33
5	other_days	1.26

tidyr::gather()

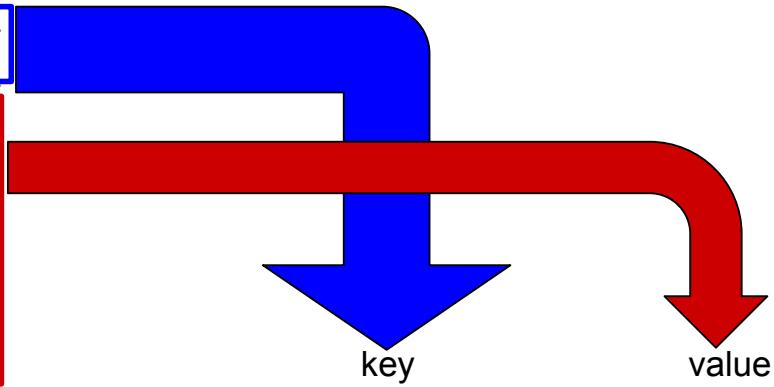
patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26



patient	day_type	mean_behavior
1	moon_days	3.33
1	other_days	0.27
2	moon_days	3.67
2	other_days	0.59
3	moon_days	2.67
3	other_days	0.32
4	moon_days	3.33
4	other_days	0.19
5	moon_days	3.33
5	other_days	1.26

tidyr::gather()

patient	moon_days	other_days
1	3.33	0.27
2	3.67	0.59
3	2.67	0.32
4	3.33	0.19
5	3.33	1.26



patient	day_type	mean_behavior
1	moon_days	3.33
1	other_days	0.27
2	moon_days	3.67
2	other_days	0.59
3	moon_days	2.67
3	other_days	0.32
4	moon_days	3.33
4	other_days	0.19
5	moon_days	3.33
5	other_days	1.26

```
tidyr::gather(data = df, key = "day_type", value = "mean_behavior", -patient)
```

tidyr::gather()

```
tidyr::gather(data = df, key = "day_type", value = "mean_behavior")
```

```
     day_type mean_behavior
1     patient          1.00
2     patient          2.00
3     patient          3.00
4     patient          4.00
5     patient          5.00
6     patient          6.00
7     patient          7.00
8     patient          8.00
9     patient          9.00
10    patient         10.00
11    patient         11.00
12    patient         12.00
13    patient         13.00
14    patient         14.00
15    patient         15.00
16  moon_days          3.33
17  moon_days          3.67
18  moon_days          2.67
19  moon_days          3.33
20  moon_days          3.33
21  moon_days          3.67
22  moon_days          4.67
23  moon_days          2.67
24  moon_days          6.00
25  moon_days          4.33
26  moon_days          3.33
27  moon_days          0.67
28  moon_days          1.33
29  moon_days          0.33
30  moon_days          2.00
31 other_days          0.27
```

Example: Is it tidy?

Mode of Delivery	Covariate			No. of Mother–Child Pairs	No. of HIV-1–Infected Children
	No. of Periods of Antiretroviral Therapy	Advanced Maternal Disease	Low Birth Weight of Infant (<2500 g)		
Elective cesarean	0	No	No	372	30
Other	0	No	No	3850	652
Elective cesarean	0	Yes	No	28	5
Other	0	Yes	No	303	74
Elective cesarean	0	No	Yes	110	17
Other	0	No	Yes	767	196
Elective cesarean	0	Yes	Yes	27	4
Other	0	Yes	Yes	114	40
Elective cesarean	1 or 2	No	No	41	0
Other	1 or 2	No	No	441	49
Elective cesarean	1 or 2	Yes	No	23	3
Other	1 or 2	Yes	No	186	33
Elective cesarean	1 or 2	No	Yes	7	0
Other	1 or 2	No	Yes	83	22
Elective cesarean	1 or 2	Yes	Yes	10	3
Other	1 or 2	Yes	Yes	54	19
Elective cesarean	3	No	No	124	2
Other	3	No	No	878	49
Elective cesarean	3	Yes	No	34	1
Other	3	Yes	No	208	24
Elective cesarean	3	No	Yes	25	0
Other	3	No	Yes	109	11
Elective cesarean	3	Yes	Yes	8	1
Other	3	Yes	Yes	38	6

Exercise: Bednets

- Model for the expected number of cases of malaria:

$$
\begin{aligned}
\mu &= \text{Rate} \times \text{Person time} \\
\mu &= \lambda \times PT \\
&= \lambda_0 \times \theta^{\text{exposed}} \times PT \\
\log(\mu) &= \log(\lambda_0) + \log(\theta) \times \text{exposed} + \log(PT)
\end{aligned}
$$

where

$$
\text{exposed} = \begin{cases} 0 & \text{standard bednet} \\ 1 & \text{treated bednet} \end{cases}
$$