Tidy Data

Journal Club May 13th, 2020

Tidy Data

Rules:

1. Each column is one variable
2. Each row is one observation
3. Each value is in one cell

Why use tidy data?

e.g. easy data visualisations in ggplot2

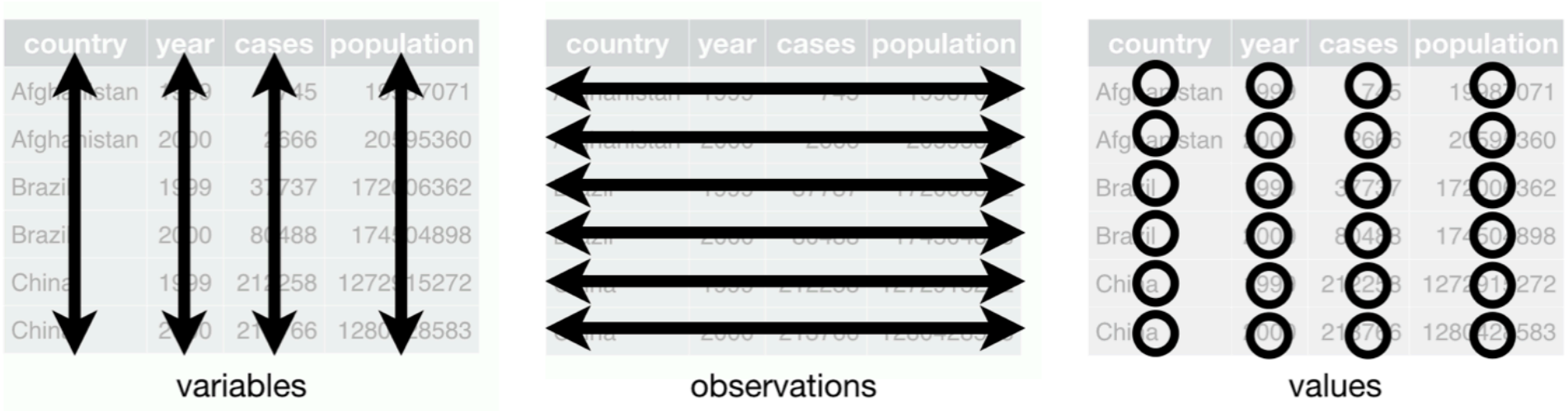
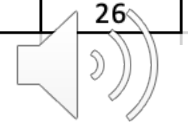


Figure 12.1: Following three rules makes a dataset tidy: variables are in columns, observations are in rows, and values are in cells.

Tidy Data

'Messy data‘

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	sampling 02.09.2019																	
2																		
3	**P1 #S3**	**quare**																
4	**spezies**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Cells**	**length_squares_400x[µm]**	**squares_area[mm2]**	**chamber_area[mm2]**	**Vol_chamber[ml]**	**Cells in 2.973ml**	**Cells/ml**
5	Leptocylindros sp.	19	15	8	10	17	15	13	8	14	16	135	250	0,625	490,874	10	10603	3566
6	Chaetoceros simplex					1	1	1			1	4	250	0,625	490,874	10	314	106
7	Dinoflagellates		1		1			1	1	1	1	6	250	0,625	490,874	10	471	159
8	Prorcendrum redfieldii								1			1	250	0,625	490,874	10	79	26
9																		
10	**P7 #S3**	**quare**																
11	**spezies**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Cells**	**length_squares_400x[µm]**	**squares_area[mm2]**	**chamber_area[mm2]**	**Vol_chamber[ml]**	**Cells in 2.973ml**	**Cells/ml**
12	Leptocylindros sp.	5	8	5	4	14	10	7	14	18	15	100	250	0,625	490,874	10	7854	2642
13	Chaetoceros simplex										1	1	250	0,625	490,874	10	79	26
14	Dinoflagellates	4		2	2			3	2	2		15	250	0,625	490,874	10	1178	396
15	Thalassiosira sp.		1									1	250	0,625	490,874	10	79	26
16	Cylindotheca sp.						1					1	250	0,625	490,874	10	79	26



Tidy Data

'Messy data‘

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	sampling 02.09.2019																	
2																		
3	**P1 #S3**	quare																
4	**spezies**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Cells**	**length_squares_400x[µm]**	**squares_area[mm2]**	**chamber_area[mm2]**	**Vol_chamber[ml]**	**Cells in 2.973ml**	**Cells/ml**
5	Leptocylindros sp.	19	15	8	10	17	15	13	8	14	16	135	250	0,625	490,874	10	10603	3566
6	Chaetoceros simplex					1	1	1			1	4	250	0,625	490,874	10	314	106
7	Dinoflagellates		1		1			1	1	1	1	6	250	0,625	490,874	10	471	159
8	Prorcendrum redfieldii								1			1	250	0,625	490,874	10	79	26
9																		
10	**P7 #S3**	quare																
11	**spezies**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Cells**	**length_squares_400x[µm]**	**squares_area[mm2]**	**chamber_area[mm2]**	**Vol_chamber[ml]**	**Cells in 2.973ml**	**Cells/ml**
12	Leptocylindros sp.	5	8	5	4	14	10	7	14	18	15	100	250	0,625	490,874	10	7854	2642
13	Chaetoceros simplex										1	1	250	0,625	490,874	10	79	26
14	Dinoflagellates	4		2	2			3	2	2		15	250	0,625	490,874	10	1178	396
15	Thalassiosira sp.		1									1	250	0,625	490,874	10	79	26
16	Cylindotheca sp.						1					1	250	0,625	490,874	10	79	26

Tidy Data

'Messy data‘

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	sampling 02.09.2019																	
2																		
3	**P1 #S3**	quare																
4	**spezies**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Cells**	**length_squares_400x[µm]**	**squares_area[mm2]**	**chamber_area[mm2]**	**Vol_chamber[ml]**	**Cells in 2.973ml**	**Cells/ml**
5	Leptocylindros sp.	19	15	8	10	17	15	13	8	14	16	135	250	0,625	490,874	10	10603	3566
6	Chaetoceros simplex					1	1	1			1	4	250	0,625	490,874	10	314	106
7	Dinoflagellates		1		1			1	1	1	1	6	250	0,625	490,874	10	471	159
8	Prorcendrum redfieldii								1			1	250	0,625	490,874	10	79	26
9																		
10	**P7 #S3**	quare																
11	**spezies**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Cells**	**length_squares_400x[µm]**	**squares_area[mm2]**	**chamber_area[mm2]**	**Vol_chamber[ml]**	**Cells in 2.973ml**	**Cells/ml**
12	Leptocylindros sp.	5	8	5	4	14	10	7	14	18	15	100	250	0,625	490,874	10	7854	2642
13	Chaetoceros simplex										1	1	250	0,625	490,874	10	79	26
14	Dinoflagellates	4		2	2			3	2	2		15	250	0,625	490,874	10	1178	396
15	Thalassiosira sp.		1									1	250	0,625	490,874	10	79	26
16	Cylindotheca sp.						1					1	250	0,625	490,874	10	79	26

Tidy Data

'Messy data‘

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	sampling 02.09.2019																	
2																		
3	**P1 #S3**	quare																
4	**spezies**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Cells**	**length_squares_400x[µm]**	**squares_area[mm2]**	**chamber_area[mm2]**	**Vol_chamber[ml]**	**Cells in 2.973ml**	**Cells/ml**
5	Leptocylindros sp.	19	15	8	10	17	15	13	8	14	16	135	250	0,625	490,874	10	10603	3566
6	Chaetoceros simplex					1	1	1			1	4	250	0,625	490,874	10	314	106
7	Dinoflagellates		1		1			1	1	1	1	6	250	0,625	490,874	10	471	159
8	Prorcendrum redfieldii								1			1	250	0,625	490,874	10	79	26
9																		
10	**P7 #S3**	quare																
11	**spezies**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Cells**	**length_squares_400x[µm]**	**squares_area[mm2]**	**chamber_area[mm2]**	**Vol_chamber[ml]**	**Cells in 2.973ml**	**Cells/ml**
12	Leptocylindros sp.	5	8	5	4	14	10	7	14	18	15	100	250	0,625	490,874	10	7854	2642
13	Chaetoceros simplex										1	1	250	0,625	490,874	10	79	26
14	Dinoflagellates	4		2	2			3	2	2		15	250	0,625	490,874	10	1178	396
15	Thalassiosira sp.		1									1	250	0,625	490,874	10	79	26
16	Cylindotheca sp.						1					1	250	0,625	490,874	10	79	26

Tidy Data

Rules:

1. Each column is one variable
2. Each row is one observation
3. Each value is in one cell

'Messy data‘

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	sampling 02.09.2019																	
2																		
3	**P1 #S3**	quare																
4	**spezies**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Cells**	**length_squares_400x[µm]**	**squares_area[mm2]**	**chamber_area[mm2]**	**Vol_chamber[ml]**	**Cells in 2.973ml**	**Cells/ml**
5	Leptocylindros sp.	19	15	8	10	17	15	13	8	14	16	135	250	0,625	490,874	10	10603	3566
6	Chaetoceros simplex					1	1	1			1	4	250	0,625	490,874	10	314	106
7	Dinoflagellates		1		1			1	1	1	1	6	250	0,625	490,874	10	471	159
8	Prorcendrum redfieldii								1			1	250	0,625	490,874	10	79	26
9																		
10	**P7 #S3**	quare																
11	**spezies**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Cells**	**length_squares_400x[µm]**	**squares_area[mm2]**	**chamber_area[mm2]**	**Vol_chamber[ml]**	**Cells in 2.973ml**	**Cells/ml**
12	Leptocylindros sp.	5	8	5	4	14	10	7	14	18	15	100	250	0,625	490,874	10	7854	2642
13	Chaetoceros simplex										1	1	250	0,625	490,874	10	79	26
14	Dinoflagellates	4		2	2			3	2	2		15	250	0,625	490,874	10	1178	396
15	Thalassiosira sp.		1									1	250	0,625	490,874	10	79	26
16	Cylindotheca sp.						1					1	250	0,625	490,874	10	79	26

Tidy Data

Rules:

1. Each column is one variable
2. Each row is one observation
3. Each value is in one cell

Variable: sampling date

Variable: sample

Empty rows no headers

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	sampling 02.09.2019																	
2																		
3	**P1 #S3**	**quare**																
4	**spezies**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Cells**	**length_squares_400x[µm]**	**squares_area[mm2]**	**chamber_area[mm2]**	**Vol_chamber[ml]**	**Cells in 2.973ml**	**Cells/ml**
5	Leptocylindros sp.	19	15	8	10	17	15	13	8	14	16	135	250	0,625	490,874	10	10603	3566
6	Chaetoceros simplex					1	1	1			1	4	250	0,625	490,874	10	314	106
7	Dinoflagellates		1		1			1	1	1	1	6	250	0,625	490,874	10	471	159
8	Prorcendrum redfieldii								1			1	250	0,625	490,874	10	79	26
9																		
10	**P7 #S3**	**quare**																
11	**spezies**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Cells**	**length_squares_400x[µm]**	**squares_area[mm2]**	**chamber_area[mm2]**	**Vol_chamber[ml]**	**Cells in 2.973ml**	**Cells/ml**
12	Leptocylindros sp.	5	8	5	4	14	10	7	14	18	15	100	250	0,625	490,874	10	7854	2642
13	Chaetoceros simplex										1	1	250	0,625	490,874	10	79	26
14	Dinoflagellates	4		2	2			3	2	2		15	250	0,625	490,874	10	1178	396
15	Thalassiosira sp.		1									1	250	0,625	490,874	10	79	26
16	Cylindotheca sp.						1					1	250	0,625	490,874	10	79	26

'Messy data'

1 square has 1 column – 1 counted sq
= 1 observation (needs to be in rows)

What to do with messy data

- Don‘t create a messy table in the first place
- Use tidyverse to gather columns and separate rows
- Change your excel (tf etc.) file to a user friendly version to save time and nerves

Create 2 new columns

Remove empty columns

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	sampling 02.09.2019																	
2																		
3	**P1 #S3**	**quare**																
4	**spezies**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Cells**	**length_squares_400x[µm]**	**squares_area[mm2]**	**chamber_area[mm2]**	**Vol_chamber[ml]**	**Cells in 2.973ml**	**Cells/ml**
5	Leptocylindros sp.	19	15	8	10	17	15	13	8	14	16	135	250	0,625	490,874	10	10603	3566
6	Chaetoceros simplex					1	1	1			1	4	250	0,625	490,874	10	314	106
7	Dinoflagellates		1		1			1	1	1	1	6	250	0,625	490,874	10	471	159
8	Prorcendrum redfieldii								1			1	250	0,625	490,874	10	79	26
9																		
10	**P7 #S3**	**quare**																
11	**spezies**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Cells**	**length_squares_400x[µm]**	**squares_area[mm2]**	**chamber_area[mm2]**	**Vol_chamber[ml]**	**Cells in 2.973ml**	**Cells/ml**
12	Leptocylindros sp.	5	8	5	4	14	10	7	14	18	15	100	250	0,625	490,874	10	7854	2642
13	Chaetoceros simplex										1	1	250	0,625	490,874	10	79	26
14	Dinoflagellates	4		2	2			3	2	2		15	250	0,625	490,874	10	1178	396
15	Thalassiosira sp.		1									1	250	0,625	490,874	10	79	26
16	Cylindotheca sp.						1					1	2	0,625	490,874	10	79	26

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	sampling	sample	**spezies**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Cells**	**length_squares_400x[µm]**	**squares_area[mm2]**	**chamber_area[mm2]**	**Vol_chamber[ml]**
2	02.09.19	**P1 #S3**	Chaetoceros simplex	19	15	8	10	17	15	13	8	14	16	135	250	0,625	490,874	10
3	02.09.19		Dinoflagellates					1	1	1			1	4	250	0,625	490,874	10
4	02.09.19		Prorcendrum redfieldii		1		1			1	1	1	1	6	250	0,625	490,874	10
5	02.09.19																	
6	02.09.19	**P7 #S3**	**spezies**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Cells**	**length_squares_400x[µm]**	**squares_area[mm2]**	**chamber_area[mm2]**	**Vol_chamber[ml]**
7	02.09.19		Leptocylindros sp.	5	8	5	4	14	10	7	14	18	15	100	250	0,625	490,874	10
8	02.09.19		Chaetoceros simplex										1	1	250	0,625	490,874	10
9	02.09.19		Dinoflagellates	4		2	2			3	2	2		15	250	0,625	490,874	10
10	02.09.19		Thalassiosira sp.		1									1	250	0,625	490,874	10
11	02.09.19		**square**															
12	02.09.19	**P10 #S3**	**spezies**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Cells**	**length_squares_400x[µm]**	**squares_area[mm2]**	**chamber_area[mm2]**	**Vol_chamber[ml]**
13	02.09.19		Leptocylindros sp.	16	11	8	13	14	9	11	12	14	10	118	250	0,625	490,874	10
14	02.09.19		Chaetoceros simplex				1				1	1		3	250	0,625	490,874	10
15	02.09.19		Dinoflagellates						2		1		1	4	250	0,625	490,874	10

Excercise

- Create a tidy data frame using either:
  - the newly created dataset examp_messydata_2
  - Or the messy data examp_messydata_1

Useful functions

- Pivot_longer()/pivot_wider() - or gather() and spread()
- Separate()
- Select()
- Filter()
- Drop_na()
- Rename()
- Slice()
- Fill()

https://github.com/rfordatascience/tidytuesday