Keeping Your Spreadsheets Tidy

Simon Esling
Centre for eResearch
University of Auckland

July 2024

Today's session

- What is Tidy Data?
- Tidy Data core principles
- Why do we want Tidy Data?
- Common mistakes with data
- Example – Messy
- Example – Tidy
- Some suggested tools

ResBaz
RESEARCH BAZAAR

What is Tidy Data?

"Tidy data is a standard way of mapping the meaning of a dataset to its structure. A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types [...] *Messy data* is any other arrangement of the data." (Wickham, p.4, 2014)

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software*, *59*(10), 1–23. https://doi.org/10.18637/jss.v059.i10

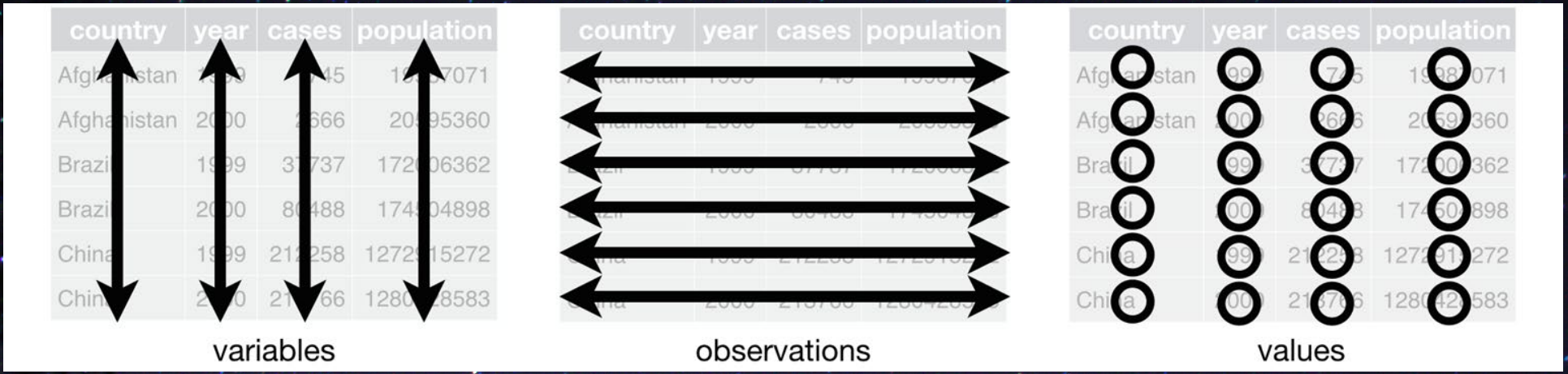
ResBaz
RESEARCH BAZAAR

What is Tidy Data?

- An approach to organising data that makes it easier to analyse it computationally.
- A lot of the things we do with spreadsheets appear to make it tidy, for us, but may present computational problems later.
- Tidy data creates a better balance between human-readable and computer-readable data.

ResBaz
RESEARCH BAZAAR

Tidy Data core principles



- Each **variable** must have its **own column**
- Each **observation** must have its **own row**
- Each **value** must have its own **cell**
- Exception: 1st row = variable names

ResBaz
RESEARCH BAZAAR

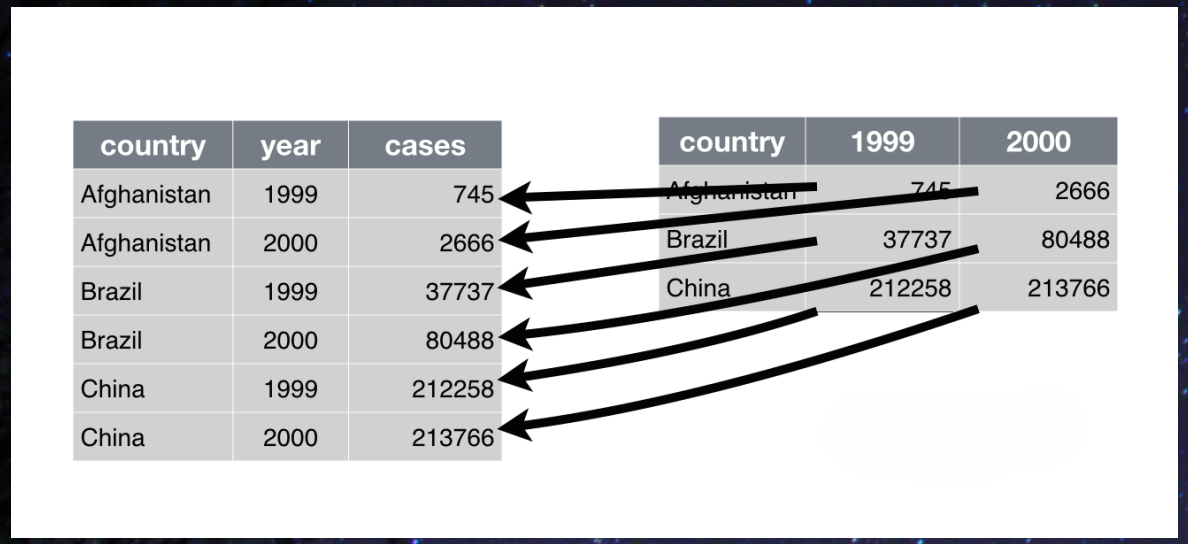
Tidy Data core principles

- Don’t combine multiple pieces of information in one cell, e.g. m25
- Don’t combine multiple tables in one Spreadsheet
- Leave the raw data raw – do not change it!
  - Make a working copy and then save new versions as needed
  - Have a sensible file naming structure
  - Keep a meta-record, e.g., README.txt file
  - Keep a record of your ‘recipe’ to tidy the data
- Export the tidy data in a CSV format (comma-separated values)
  - CSV is a plain text, open format that is more future-proof
  - CSV is a format requirement by most data repositories

ResBaz
RESEARCH BAZAAR

Tidy Data core principles

The *longer* format



country	year	cases
Afghanistan	1999	745
Afghanistan	2000	2666
Brazil	1999	37737
Brazil	2000	80488
China	1999	212258
China	2000	213766

country	1999	2000
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

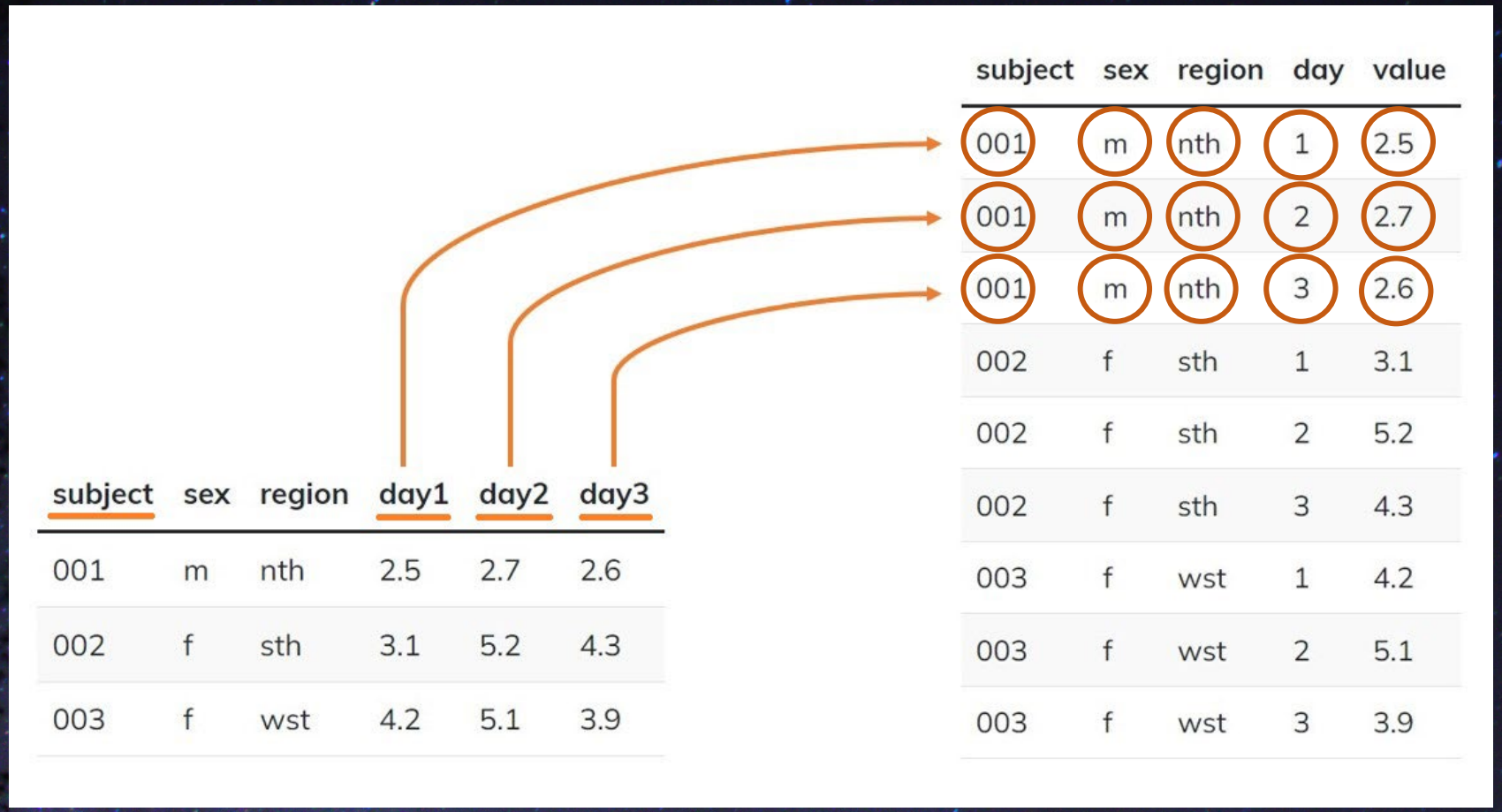
The *wider* format

country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

ResBaz
RESEARCH BAZAAR

Tidy Data core principles

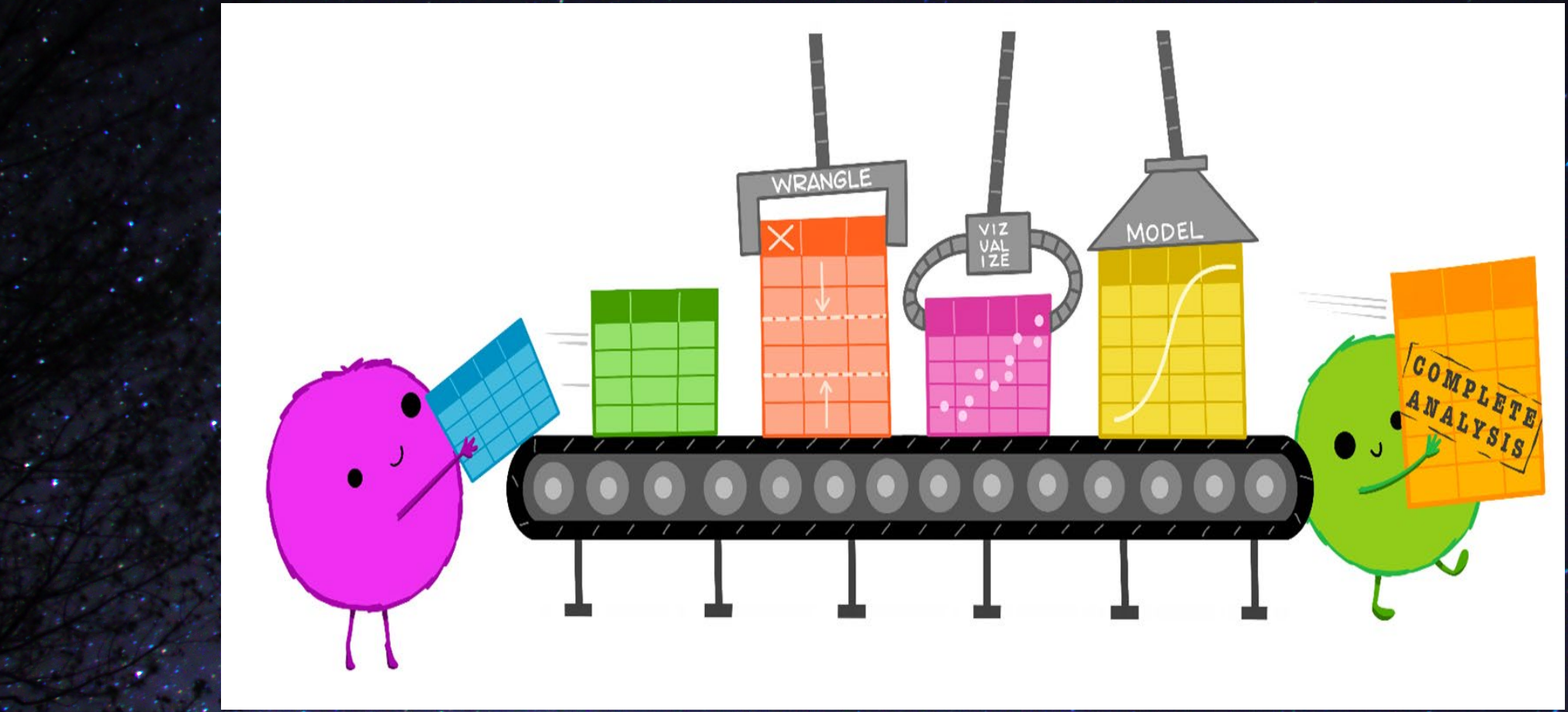


subject	sex	region	day1	day2	day3
001	m	nth	2.5	2.7	2.6
002	f	sth	3.1	5.2	4.3
003	f	wst	4.2	5.1	3.9

subject	sex	region	day	value
001	m	nth	1	2.5
001	m	nth	2	2.7
001	m	nth	3	2.6
002	f	sth	1	3.1
002	f	sth	2	5.2
002	f	sth	3	4.3
003	f	wst	1	4.2
003	f	wst	2	5.1
003	f	wst	3	3.9

ResBaz
RESEARCH BAZAAR

Why do we want Tidy Data?



ResBaz
RESEARCH BAZAAR

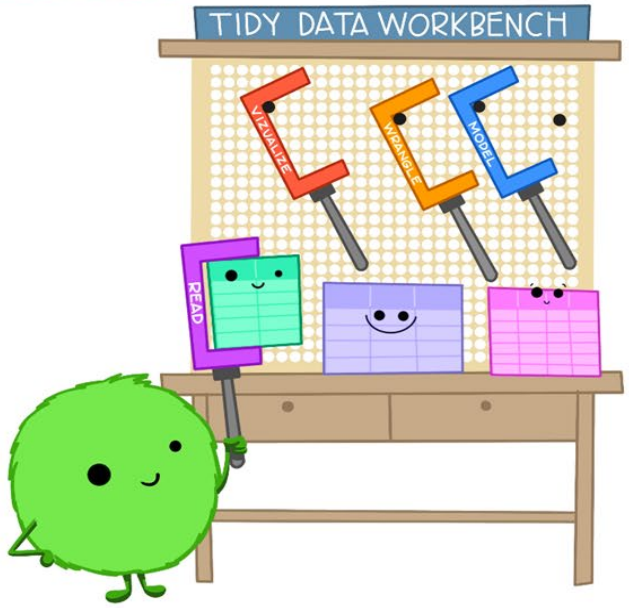
Why do we want Tidy Data?

- Get better insights from the data, easier to work with, when it is tidy
  - Find, consolidate and reconcile errors in the data
  - Reproducibility easier for yourself and other researchers
- Tidy data plays well with programming languages – R, Python
  - Lateral thinking vs. computational power
  - Machine-readable → ease of finding and processing data
  - Code can be re-run on future datasets
- Publisher requirement (showing all the steps of analysis)
- Frontloading → time invested now, saves more time in the future

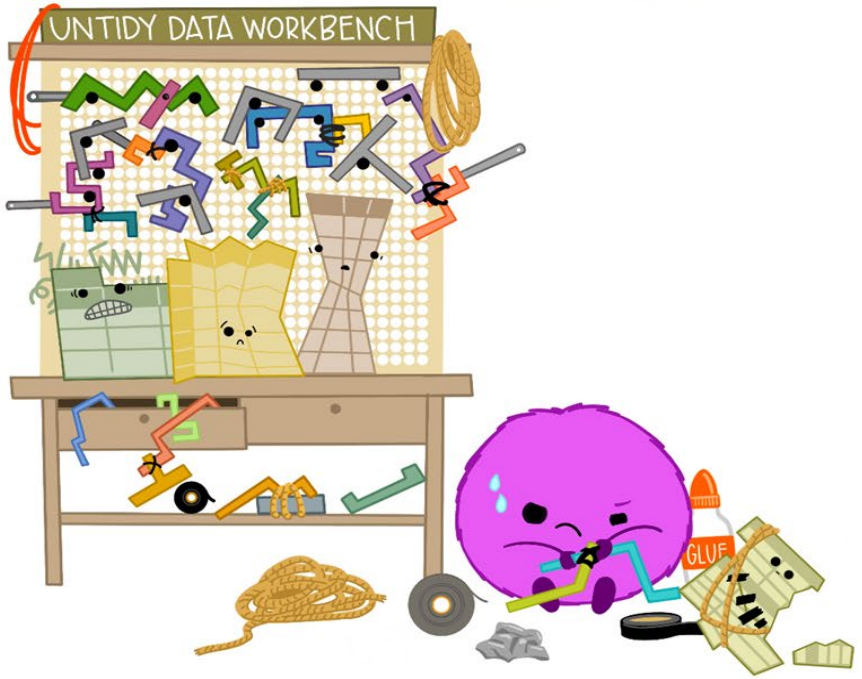
ResBaz
RESEARCH BAZAAR

Why do we want Tidy Data?

When working with tidy data, we can use the same tools in similar ways for different datasets...



...but working with untidy data often means reinventing the wheel with one-time approaches that are hard to iterate or reuse.



ResBaz
RESEARCH BAZAAR

Common mistakes with data

- Treating spreadsheet programs like lab notebooks, relying on context, notes in the margin, spatial layout and fields to convey information
- Computers do not interpret information the way we do and will not ‘see’ how our data fits together when there is erroneous information
- Using special characters (/ $ % # @ \ -
- Inconsistent column names, e.g., ‘sample_a’ or ‘sample_A’
- Leading/trailing whitespace, e.g., ‘ sample_a’ or ‘sample_a ’

ResBaz
RESEARCH BAZAAR

Common mistakes with data

- Merging cells and other formatting for aesthetic reasons
- Highlighting and formatting within the cells
- More than one piece of information per cell
- Having different tables of data in the same sheet
- Not recording zeros as zeros
- Using different null values to indicate missing data
- Date formatting inconsistencies, e.g., 10/2/2023, 2023-02-10 or 10_2_23

ResBaz
RESEARCH BAZAAR

Example – Messy

2013 Field Season

Species: DM			
Date Collected	Plot	Sex	Weight
16/07/2013	2	F	
16/07/2013	7	M	33g
16/07/2013	3	M	
16/07/2013	1	M	
18/07/2013	3	M	40g
18/07/2013	7	M	48g
18/07/2013	4	F	29g
18/07/2013	4	F	46g
18/07/2013	7	M	36g
18/07/2013	7	F	35g
18/07/2013	8	F	22g
18/07/2013	7	F	42g
18/07/2013	4	F	41g
18/07/2013	6	F	37g

Species: DO			
Date Collected	Plot	Sex	Weight
19/08/2013	8	F	52
17/10/2013	3	F	33
17/10/2013	3	F	50
17/10/2013	17	F	48
17/10/2013	17	F	31
18/10/2013	8	F	41
12/11/2013	1	F	44
12/11/2013	1	M	48
14/11/2013	8	F	39
10/12/2013	9	F	40
10/12/2013	1	M	45
11/12/2013	8	F	41

Species: DS			
Date Collected	Plot	Sex	Weight
12/11/2013	9	F	117
12/11/2013	1	F	121
12/11/2013	20	M	115
12/11/2013	9	F	120
13/11/2013	17	F	118
13/11/2013	11	F	126
13/11/2013	17	M	132 (scale not calibrated)
13/11/2013	14	F	113 (scale not callibtrated)
13/11/2013	11	F	122
13/11/2013	4	F	107
13/11/2013	4	F	115

ResBaz
RESEARCH BAZAAR

Example – Messy

2013 Field Season

Species: DM

Date Collected	Plot	Sex	Weight
16/07/2013	2	F	
16/07/2013	7	M	33g
16/07/2013	3	M	
16/07/2013	1	M	
18/07/2013	3	M	40g
18/07/2013	7	M	48g
18/07/2013	4	F	29g
18/07/2013	4	F	46g
18/07/2013	7	M	36g
18/07/2013	7	F	35g
18/07/2013	8	F	22g
18/07/2013	7	F	42g
18/07/2013	4	F	41g
18/07/2013	6	F	37g

Species: DO

Date Collected	Plot	Sex	Weight
19/08/2013	8	F	52
17/10/2013	3	F	33
17/10/2013	3	F	50
17/10/2013	17	F	48
17/10/2013	17	F	31
18/10/2013	8	F	41
12/11/2013	1	F	44
12/11/2013	1	M	48
14/11/2013	8	F	39
10/12/2013	9	F	40
10/12/2013	1	M	45
11/12/2013	8	F	41

Species: DS

Date Collected	Plot	Sex	Weight
12/11/2013	9	F	117
12/11/2013	1	F	121
12/11/2013	20	M	115
12/11/2013	9	F	120
13/11/2013	17	F	118
13/11/2013	11	F	126
13/11/2013	17	M	132 (scale not calibrated)
13/11/2013	14	F	113 (scale not callibtrated)
13/11/2013	11	F	122
13/11/2013	4	F	107
13/11/2013	4	F	115

ResBaz
RESEARCH BAZAAR

Example – Messy

Plot: 1			
Date collected	Species	Sex	Weight
9/01/2014	DM	M	40
9/01/2014	DM	F	36
9/01/2014	DS	F	135
20/01/2014	DM	F	39
20/01/2014	DM	M	43
20/01/2014	DS	F	144
13/03/2014	DM	F	51
13/03/2014	DM	F	44
13/03/2014	DS	F	146

Plot: 2			
Date collected	Species	Sex	Weight
8/01/2014	NA		
8/01/2014	DM	M	44
8/01/2014	DM	M	38
8/01/2014	OL		
8/01/2014	PE	M	22
8/01/2014	DM	M	38
8/01/2014	DM	M	48
8/01/2014	DM	M	43
8/01/2014	DM	F	35
8/01/2014	DM	M	43
8/01/2014	DM	F	37
8/01/2014	PF	F	7
8/01/2014	DM	M	45
8/01/2014	OT		
8/01/2014	DS	M	157
8/01/2014	OX		
18/02/2014	NA	M	218
18/02/2014	PF	F	7
18/02/2014	DM	M	52

Plot: 3			
Date collected	Species	Sex	Weight
1/8	PF	M	7
2/18	OT	M	24
2/18	OT	F	23
3/11	NA	M	232
3/11	OT	F	22
3/11	OT	M	26
3/11	PF	M	8
4/8	NA	F	
5/6			
5/18	NA	F	182
6/9	OT	F	29
7/8	NA	F	115
7/8	NA	M	190

Plot: 4		
Date collected	species_sex	wgt
8/01/1978	DM_F	37
8/01/1978	DS_F	128
8/01/1978	DM_F	42
8/01/1978	DM_M	37
8/01/1978	DM_M	
8/01/1978	DM_F	48
8/01/1978	DM_M	45
8/01/1978	DM_F	42
8/01/1978	DO_M	52
8/01/1978	OL_M	35

gray cell means my measurement device wasn't calibrated correctly

ResBaz
RESEARCH BAZAAR

Example – Messy

Plot: 1

Date collected	Species	Sex	Weight
9/01/2014	DM	M	40
9/01/2014	DM	F	36
9/01/2014	DS	F	135
20/01/2014	DM	F	39
20/01/2014	DM	M	43
20/01/2014	DS	F	144
13/03/2014	DM	F	51
13/03/2014	DM	F	44
13/03/2014	DS	F	146

Plot: 2

Date collected	Species	Sex	Weight
8/01/2014	NA		
8/01/2014	DM	M	44
8/01/2014	DM	M	38
8/01/2014	OL		
8/01/2014	PE	M	22
8/01/2014	DM	M	38
8/01/2014	DM	M	48
8/01/2014	DM	M	43
8/01/2014	DM	F	35
8/01/2014	DM	M	43
8/01/2014	DM	F	37
8/01/2014	PF	F	7
8/01/2014	DM	M	45
8/01/2014	OT		
8/01/2014	DS	M	157
8/01/2014	OX		
18/02/2014	NA	M	218
18/02/2014	PF	F	7
18/02/2014	DM	M	52

Plot: 3

Date collected	Species	Sex	Weight
1/8	PF	M	7
2/18	OT	M	24
2/18	OT	F	23
3/11	NA	M	232
3/11	OT	F	22
3/11	OT	M	26
3/11	PF	M	8
4/8	NA	F	
5/6			
5/18	NA	F	182
6/9	OT	F	29
7/8	NA	F	115
7/8	NA	M	190

Plot: 4

Date collected	species_sex	wgt
8/01/1978	DM_F	37
8/01/1978	DS_F	128
8/01/1978	DM_F	42
8/01/1978	DM_M	37
8/01/1978	DM_M	
8/01/1978	DM_F	48
8/01/1978	DM_M	45
8/01/1978	DM_F	42
8/01/1978	DO_M	52
8/01/1978	OL_M	35

gray cell means my measurement device wasn't calibrated correctly

ResBaz
RESEARCH BAZAAR

Example – Tidy

	A	B	C	D	E	F	G	H	I
1	year	month	day	species	plot	sex	weight	calibrated	
26	2013	12	10	DO	1	M	45		
27	2013	12	11	DO	8	F	41		
28	2013	11	12	DS	9	F	117		
29	2013	11	12	DS	1	F	121		
30	2013	11	12	DS	20	M	115		
31	2013	11	12	DS	9	F	120		
32	2013	11	13	DS	17	F	118		
33	2013	11	13	DS	11	F	126		
34	2013	11	13	DS	17	M	132	NO	
35	2013	11	13	DS	14	F	113	NO	
36	2013	11	13	DS	11	F	122		
37	2013	11	13	DS	4	F	107		
38	2013	11	13	DS	4	F	115		
39	2014	1	9	DM	1	M	40		
40	2014	1	9	DM	1	F	36		
41	2014	1	9	DS	1	F	135		
42	2014	1	20	DM	1	F	39		
43	2014	1	20	DM	1	M	43		
44	2014	1	20	DS	1	F	144		
45	2014	3	13	DM	1	F	51		

ResBaz
RESEARCH BAZAAR

Example – Tidy

	A	B	C	D	E	F	G	H	I
1	year	month	day	species	plot	sex	weight	calibrated	
26	2013	12	10	DO	1	M	45		
27	2013	12	11	DO	8	F	41		
28	2013	11	12	DS	9	F	117		
29	2013	11	12	DS	1	F	121		
30	2013	11	12	DS	20	M	115		
31	2013	11	12	DS	9	F	120		
32	2013	11	13	DS	17	F	118		
33	2013	11	13	DS	11	F	126		
34	2013	11	13	DS	17	M	132	NO	
35	2013	11	13	DS	14	F	113	NO	
36	2013	11	13	DS	11	F	122		
37	2013	11	13	DS	4	F	107		
38	2013	11	13	DS	4	F	115		
39	2014	1	9	DM	1	M	40		
40	2014	1	9	DM	1	F	36		
41	2014	1	9	DS	1	F	135		
42	2014	1	20	DM	1	F	39		
43	2014	1	20	DM	1	M	43		
44	2014	1	20	DS	1	F	144		
45	2014	3	13	DM	1	F	51		

ResBaz
RESEARCH BAZAAR

Some suggested tools

OpenRefine – https://openrefine.org/

R – https://www.r-project.org/about.html

Python – https://www.python.org/about/


Questions?