Cogs 9
Discussion Section

FA22 Week 4
Will McCarthy

Upcoming due dates

Quiz 3: Thursday 20th Oct

Assignment 2: Friday, October 28th

Mid Way Team Evaluations E.C: Friday, October 28th

Grades back to you soon- some gradescope issues on our end, sorry!

This week’s content

3(b): Broman & Woo, 2018, Data organization in spreadsheets (IAs)

3(a): Hadley Wickham, 2014, Tidy Data

- Data definitions
- What is Tidy data?
- Melting

Data wrangling demo

Reading 3(a): Hadley Wickham, 2014, Tidy Data

80% of data analysis is spent on the process of cleaning and preparing the data (Dasu and Johnson 2003).

Good wrangling makes solving the problem easy.

Data comes in many forms— standardization makes things easier.

Structure and Semantics

Structure is a datasets physical **layout**

Semantics is the **meaning** of its elements (columns, rows, cells)

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

Table 1: Typical presentation dataset.

These include the same data, but structured differently.

Semantics of the *elements* are the same.

The semantics of the *rows* and *columns* are different

=> ‘rows’ and ‘columns’ are arbitrary?

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

Table 2: The same data as in Table 1 but structured differently.

Useful semantic categories: *Variables*, *Values*, *Observations*

Difference between a variable and a value?

Variables can take many *values*

Observations are a single recorded unit

Eg. An experimental trial, a recording, a measurement, an event

Different *kinds* of observations might exist within the same dataset (e.g. a trial and an exit survey response)

Variables

Observations

Values

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

Values

Values

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

What is Tidy Data?

“Tidy” doesn’t just mean neat

“Tidy” is a **framework** for organizing data

If all data share the same framework => easier to work with

  “Tidy” is the closest we have to an industry standard of data organization

Data science software packages are often designed to work with tidy data

  e.g. visualization packages (more next week)

It is a *choice* you can make about organizing your data (and often a good one)

Three characteristics of a tidy dataset

1. Each variable forms a column
2. Each observation forms a row
3. Each type of observational unit forms a table

Which of these is tidy?

Neither!

This one is tidy:

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

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

Table 1: Typical presentation dataset.

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

Table 2: The same data as in Table 1 but structured differently.

Now you try: design a tidy table

1. Each variable forms a column
2. Each observation forms a row
3. Each type of observational unit forms a table

Say you are running a psychology study. Each participant will perform 10 trials, and you need to store the response time for each trial.

Design a tidy scheme for storing this data.

(participant id, trial number, response time)

What if 5 of these trials were in condition A and five were in Condition B?

Now you try 2 (harder)

1. Each variable forms a column
2. Each observation forms a row
3. Each type of observational unit forms a table

Say you were running a psychology study measuring change in performance on some cognitive task over time. Participants will perform the task 5 times across several weeks, and you need to record their score each time to see if it improves. In weeks 1, 3, and 5 they will also take a survey, for which you need to save their responses (3 questions, each a text response).

Design a tidy scheme for storing this data.

Tidying messy datasets

Wickham steps through five common problems with messy datasets.

We'll go through the first together

- Column headers are values, not variable names.
- Multiple variables are stored in one column.
- Variables are stored in both rows and columns.
- Multiple types of observational units are stored in the same table.
- A single observational unit is stored in multiple tables.

A messy table

religion	<$10k	$10–20k	$20–30k	$30–40k	$40–50k	$50–75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

Table 4: The first ten rows of data on income and religion from the Pew Forum. Three columns, $75–100k, $100–150k and >150k, have been omitted.

Why is this not tidy?

Columns are *values,* not variables

Nothing inherently wrong with this!

(Great for presenting data)

But *not tidy*

Removing values from columns: *melting / stacking*

Variable Values

row	a	b	c
A	1	4	7
B	2	5	8
C	3	6	9

(a) Raw data

row	column	value
A	a	1
B	a	2
C	a	3
A	b	4
B	b	5
C	b	6
A	c	7
B	c	8
C	c	9

(b) Molten data

Turn columns into rows. How?

1. Keep the columns that are variables

2. Create two new columns: *column* and *value*

3. For each value column, create a stack of new rows and add the column name to the *column* column, and the values to the *value* column

Note: this is *less efficient* from a data storage standpoint (because of redundancy)

Removing values from columns: *melting / stacking*

Weeks →

year	artist	track	time	date.entered	wk1	wk2	wk3
2000	2 Pac	Baby Don't Cry	4:22	2000-02-26	87	82	72
2000	2Ge+her	The Hardest Part Of ...	3:15	2000-09-02	91	87	92
2000	3 Doors Down	Kryptonite	3:53	2000-04-08	81	70	68
2000	98^0	Give Me Just One Nig...	3:24	2000-08-19	51	39	34
2000	A*Teens	Dancing Queen	3:44	2000-07-08	97	97	96
2000	Aaliyah	I Don't Wanna	4:15	2000-01-29	84	62	51
2000	Aaliyah	Try Again	4:03	2000-03-18	59	53	38
2000	Adams, Yolanda	Open My Heart	5:30	2000-08-26	76	76	74

Table 7: The first eight Billboard top hits for 2000. Other columns not shown are `wk4`, `wk5`, ..., `wk75`.

Weeks ↓

year	artist	time	track	date	week	rank
2000	2 Pac	4:22	Baby Don't Cry	2000-02-26	1	87
2000	2 Pac	4:22	Baby Don't Cry	2000-03-04	2	82
2000	2 Pac	4:22	Baby Don't Cry	2000-03-11	3	72
2000	2 Pac	4:22	Baby Don't Cry	2000-03-18	4	77
2000	2 Pac	4:22	Baby Don't Cry	2000-03-25	5	87
2000	2 Pac	4:22	Baby Don't Cry	2000-04-01	6	94
2000	2 Pac	4:22	Baby Don't Cry	2000-04-08	7	99
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-02	1	91
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-09	2	87
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-16	3	92
2000	3 Doors Down	3:53	Kryptonite	2000-04-08	1	81
2000	3 Doors Down	3:53	Kryptonite	2000-04-15	2	70
2000	3 Doors Down	3:53	Kryptonite	2000-04-22	3	68
2000	3 Doors Down	3:53	Kryptonite	2000-04-29	4	67
2000	3 Doors Down	3:53	Kryptonite	2000-05-06	5	66

Table 8: First fifteen rows of the tidied Billboard dataset. The `date` column does not appear in the original table, but can be computed from `date.entered` and `week`.

Tidy is not about being more efficient or compressed.

Tidy is:

1. Each variable forms a column
2. Each observation forms a row
3. Each type of observational unit forms a table

Demo

Group work / questions

Future Readings

4(a): Evan M. Peck, et al., 2019, Attitudes and Perceptions of Data Visualization

4(b): Hadley Wickham, et al., 2010, Graphical Inference for Infovis

5(a): Nicholas Diakopoulos, 2016, Accountability in Algorithmic Decision Making

5(b): Julia Angwin, et al., 2016, Machine Bias