Week 04

Tidying & Tables

INFO 3402: Information Exposition

Brian Keegan, Ph.D.
Assistant Professor, Information Science
brian.keegan@colorado.edu

University of Colorado Boulder

Course Overview

Module	Week	Dates	Computational skill	Communication skill
Shaping	1	Jan 11, Jan 13	Loading	Documentation
	2	Jan 18, Jan 20	Aggregation	Summarization
	3	Jan 25, Jan 27	Joining	Validation
	4	Feb 1, Feb 3	Tidying	Tables
Distribution	5	Feb 8, Feb 10	Histograms	Perception
	6	Feb 15, Feb 17	Box plots	Audience
Comparison	7	Feb 22, Feb 24	Cat plots	Objectives
	8	Mar 1, Mar 3	Faceted plots	Simplicity
Trend	9	Mar 8, Mar 10	Line plots	Trust
	10	Mar 15, Mar 17	Stacked plots	Annotation
	11	Mar 22, Mar 24	Spring Break	
Relationship	12	Mar 29, Mar 31	Scatter plots	Fallacies
	13	Apr 5, Apr 7	Heatmaps	Persuasion
Spatial	14	Apr 12, Apr 14	Choropleths	Conventions
	15	Apr 19, Apr 21	Point plots	Design
Projects	16	Apr 26, Apr 28	Projects	

Recap – Module 01

Week	Skills	Datasets
01	Data science mindset; loading data; documentation	Census; Boulder+Broomfield weather
02	Types of data; groupby-aggregation; pivot tables	CDC deaths; time use
03	Concatenation; joining/merging	Baseball; county cannabis, crime, population, & COVID
04	Tidying data; melting/unmalting; table formatting	DIA and CDOT traffic
Module Assignment 01 due on Tuesday, February 08 by 11am		

Weekly overview

- Notebook exercises (ungraded) – Assigned Tuesdays and reviewed Thursdays
- Add a Visualization Critique (ungraded) – Discussed in class on Thursdays
- Complete "Thursday Questions" form! (ungraded/optional) – Review in class on Thursdays
- Weekly Assignment (graded, 2%) – Assigned Tuesdays and due Fridays*
- Weekly Quiz (graded, 1%) – End of class on Thursdays (open 12:00pm – 12:30pm)
- Module Assignment 01 (graded, 5%) – Due Tuesday, February 8 by 11am*

* We'll discuss due dates in a few slides

Readings

- Questions for Friday's Weekly Quiz 04 will be drawn from these readings

- pandas User Guide - Reshaping by stacking and unstacking (for WA04)
- pandas User Guide - Reshaping by melt (for WA04)
- Peng & Matsui (2018) – Chapters 1,3,10 (for MA01) → Short, easy-to-read chapters
  - Chapter 1: "Data Analysis as Art"
  - Chapter 3: "Stating and Refining the Question"
  - Chapter 10: "Communication"
- Abela (2013) – Ch. 6 – "Sequencing Your Evidence" (for MA01) → Short, easy-to-read chapters
- Optional reading: Wickham, H. (2014). "Tidy Data."

Weekly Assignment 04

- **Skills**: Reviewing previous skills plus reshaping data with melt/stack and pivot/unstack
- **Data**: CDOT traffic counters 2019 through 2021

- Lighter so you have more time for Module Assignment 01 due on Tuesday, February 8 (?)

- Lots of requests for extensions on WA03 – is Friday by midnight still a good deadline?
  - Complete this form before Thursday: https://forms.gle/MTjCbwFdmAsKHZA59
  - We'll share results and discuss options on Thursday

- Due Friday (?) before midnight on Canvas
  - Save an HTML version of your notebook with all output present and upload the HTML file to Canvas

Module Assignment 01

Module Assignment 01

- Tell a story using any of the datasets from Weeks 01 – 04
  - Census; Weather; CDC deaths; Time use; Baseball; County cannabis/crime/COVID; DIA/CDOT traffic
- 700 – 1000 words with at least one table and one visualization
- Use Peng & Matsui (2018) and Abela (2013) readings on forming a question, structuring a story
- Module Assignment 01 will be due Tuesday, February 08 by 11am
  - Submit URL of your Medium post to Canvas or save and submit as an HTML file
  - Tag your post on Medium with “INFO3402S22MA1” and whatever other tags you’d like

Using Medium

- We will be using Medium, a popular blogging platform launched in 2012
- A WYSIWYG interface like WordPress, upvoting like Reddit, tags like Instagram/Tumblr/TikTok
- Create account using your Colorado.edu email address and no need to purchase membership
- Posts will be submitted to our class publication and tag
  - https://medium.com/information-expositions-s2022 and Info3402s22
- Use the Medium Help Center or Google around for tips and advice on writing Medium posts

- Submit your Medium user name here

Rubric

- Tell a story using any of the data from Weeks 01 – 04

- Do you have a persuasive motivation?
- Are you asking a “Good Question”? (Peng & Matsui Chapter 3)
- Is your analysis appropriate for the data and your question?
- Is your evidence well-sequenced to tell a compelling story? (Abela Chapter 6)
- Does your post have one table and one visualization?
- Do your evidence and storytelling answer your question?
- Do you use a professional tone, style, and grammar with few glaring errors?

- Benchmarking to help us identify where and how to prioritize future coursework and assignments

Objections

- “We haven’t even covered how to use Medium or write a post!”
  - We’ll spend time on this on Thursday, but start to explore it yourself!
- “This feels too open-ended – just tell me what to do step-by-step like the weekly assignments!”
  - That’s not how real-world projects work!
- ”Having WA04 due on Friday and MA01 due on Tuesday is a lot of work!”
  - WA04 is lighter than previous weekly assignments
  - 3 credit-hour lecture should be a minimum of 300 minutes out-of-class work per week
  - I remain open to suggestions for alternative due dates for WAs and MAs
- “I don’t want to write on Medium”
  - Medium posts/portfolio have had **significant** benefits for previous 3402 students on job/internship market
  - See ”Publishing” section on Syllabus (pg. 3): “If students are unable or do not want to use the Medium platform, they should email the instructor before Friday, January 21 to work out an alternative arrangement.”
  - You’re welcome to write under a pseudonym as long as it’s mapped via the Account linking form

Tidy data

Tidy data

- Each variable forms a column
- Each observation forms a row
- Each observational unit forms a table

- More of a design philosophy that's widely-shared than always superior choice
  - Converting between wide and tidy is crucial!
  - Wide data is superior for many use cases

- Tidy data is *especially* important as we move into visualization for the rest of the class!

"Wide"

patient	treatment_a	treatment_b
John Smith	n/a	2
Jane Doe	16	11
Mary Johnson	3	1

"Tidy"

patient	treatment	result
John Smith	a	n/a
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Wickham (2014). "Tidy Data." *Journal of Statistical Software*.

Common problems with data

1. Column headers are values, not variable names
2. Multiple variables stored in one column
3. Variables are stored in both rows and columns
4. Multiple observational units stored in the same table
5. A single observational unit is stored in multiple tables

Column headers are values, not variable names

- 3 variables before:
  - Religion (rows)
  - Income (columns)
  - Frequency (cells)
- Melt/stack → convert columns into rows
- 3 variables after, 1 per column:
  - Religion, Income, Frequency

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

religion	income	freq
Agnostic	<$10k	27
Agnostic	$10-20k	34
Agnostic	$20-30k	60
Agnostic	$30-40k	81
Agnostic	$40-50k	76
Agnostic	$50-75k	137
Agnostic	$75-100k	122
Agnostic	$100-150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

Multiple variables stored in one column

country	year	m014	m1524	m2534	m3544	m4554	m5564	m65	mu	f014
AD	2000	0	0	1	0	0	0	0	—	—
AE	2000	2	4	4	6	5	12	10	—	3
AF	2000	52	228	183	149	129	94	80	—	93
AG	2000	0	0	0	0	0	0	1	—	1
AL	2000	2	19	21	14	24	19	16	—	3
AM	2000	2	152	130	131	63	26	21	—	1
AN	2000	0	0	1	2	0	0	0	—	0
AO	2000	186	999	1003	912	482	312	194	—	247
AR	2000	97	278	594	402	419	368	330	—	121
AS	2000	—	—	—	—	1	1	—	—	—

country	year	column	cases
AD	2000	m014	0
AD	2000	m1524	0
AD	2000	m2534	1
AD	2000	m3544	0
AD	2000	m4554	0
AD	2000	m5564	0
AD	2000	m65	0
AE	2000	m014	2
AE	2000	m1524	4
AE	2000	m2534	4
AE	2000	m3544	6
AE	2000	m4554	5
AE	2000	m5564	12
AE	2000	m65	10
AE	2000	f014	3

(a) Molten data

country	year	sex	age	cases
AD	2000	m	0-14	0
AD	2000	m	15-24	0
AD	2000	m	25-34	1
AD	2000	m	35-44	0
AD	2000	m	45-54	0
AD	2000	m	55-64	0
AD	2000	m	65+	0
AE	2000	m	0-14	2
AE	2000	m	15-24	4
AE	2000	m	25-34	4
AE	2000	m	35-44	6
AE	2000	m	45-54	5
AE	2000	m	55-64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

(b) Tidy data

Variables stored in both rows and columns

id	year	month	element	d1	d2	d3	d4	d5	d6	d7	d8
MX17004	2010	1	tmax	—	—	—	—	—	—	—	—
MX17004	2010	1	tmin	—	—	—	—	—	—	—	—
MX17004	2010	2	tmax	—	27.3	24.1	—	—	—	—	—
MX17004	2010	2	tmin	—	14.4	14.4	—	—	—	—	—
MX17004	2010	3	tmax	—	—	—	—	32.1	—	—	—
MX17004	2010	3	tmin	—	—	—	—	14.2	—	—	—
MX17004	2010	4	tmax	—	—	—	—	—	—	—	—
MX17004	2010	4	tmin	—	—	—	—	—	—	—	—
MX17004	2010	5	tmax	—	—	—	—	—	—	—	—
MX17004	2010	5	tmin	—	—	—	—	—	—	—	—

id	date	element	value
MX17004	2010-01-30	tmax	27.8
MX17004	2010-01-30	tmin	14.5
MX17004	2010-02-02	tmax	27.3
MX17004	2010-02-02	tmin	14.4
MX17004	2010-02-03	tmax	24.1
MX17004	2010-02-03	tmin	14.4
MX17004	2010-02-11	tmax	29.7
MX17004	2010-02-11	tmin	13.4
MX17004	2010-02-23	tmax	29.9
MX17004	2010-02-23	tmin	10.7

(a) Molten data

id	date	tmax	tmin
MX17004	2010-01-30	27.8	14.5
MX17004	2010-02-02	27.3	14.4
MX17004	2010-02-03	24.1	14.4
MX17004	2010-02-11	29.7	13.4
MX17004	2010-02-23	29.9	10.7
MX17004	2010-03-05	32.1	14.2
MX17004	2010-03-10	34.5	16.8
MX17004	2010-03-16	31.1	17.6
MX17004	2010-04-27	36.3	16.7
MX17004	2010-05-27	33.2	18.2

(b) Tidy data

Multiples types in one table

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

id	artist	track	time
1	2 Pac	Baby Don't Cry	4:22
2	2Ge+her	The Hardest Part Of ...	3:15
3	3 Doors Down	Kryptonite	3:53
4	3 Doors Down	Loser	4:24
5	504 Boyz	Wobble Wobble	3:35
6	98ˆ0	Give Me Just One Nig...	3:24
7	A*Teens	Dancing Queen	3:44
8	Aaliyah	I Don't Wanna	4:15
9	Aaliyah	Try Again	4:03
10	Adams, Yolanda	Open My Heart	5:30
11	Adkins, Trace	More	3:05
12	Aguilera, Christina	Come On Over Baby	3:38
13	Aguilera, Christina	I Turn To You	4:00
14	Aguilera, Christina	What A Girl Wants	3:18
15	Alice Deejay	Better Off Alone	6:50

id	date	rank
1	2000-02-26	87
1	2000-03-04	82
1	2000-03-11	72
1	2000-03-18	77
1	2000-03-25	87
1	2000-04-01	94
1	2000-04-08	99
2	2000-09-02	91
2	2000-09-09	87
2	2000-09-16	92
3	2000-04-08	81
3	2000-04-15	70
3	2000-04-22	68
3	2000-04-29	67
3	2000-05-06	66

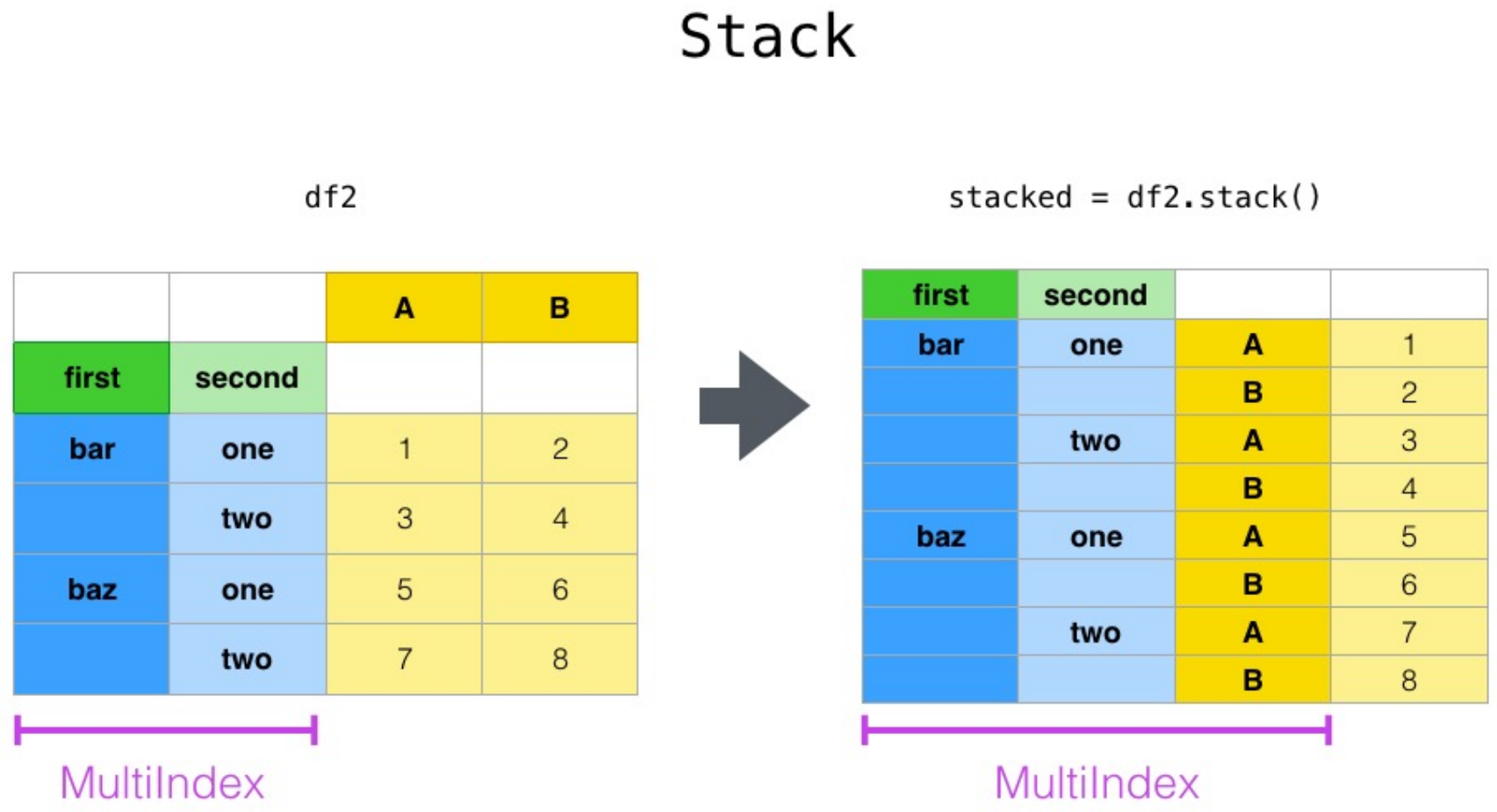
Making data tidy

- id variables are the columns in the wide data we want to keep as rows for each observation
- value variables are the columns in the wide data containing the observations to turn into rows

- Wide data to tidy data: `pd.melt` or `df.stack` (with `df.set_index()` and `df.reset_index()`)
  - **Melt strategy**: Pass id variables and/or value variables to parameters in `pd.melt`
  - **Stack strategy**: `.set_index()` on id variables, `.stack()` remaining columns, then `.reset_index()`

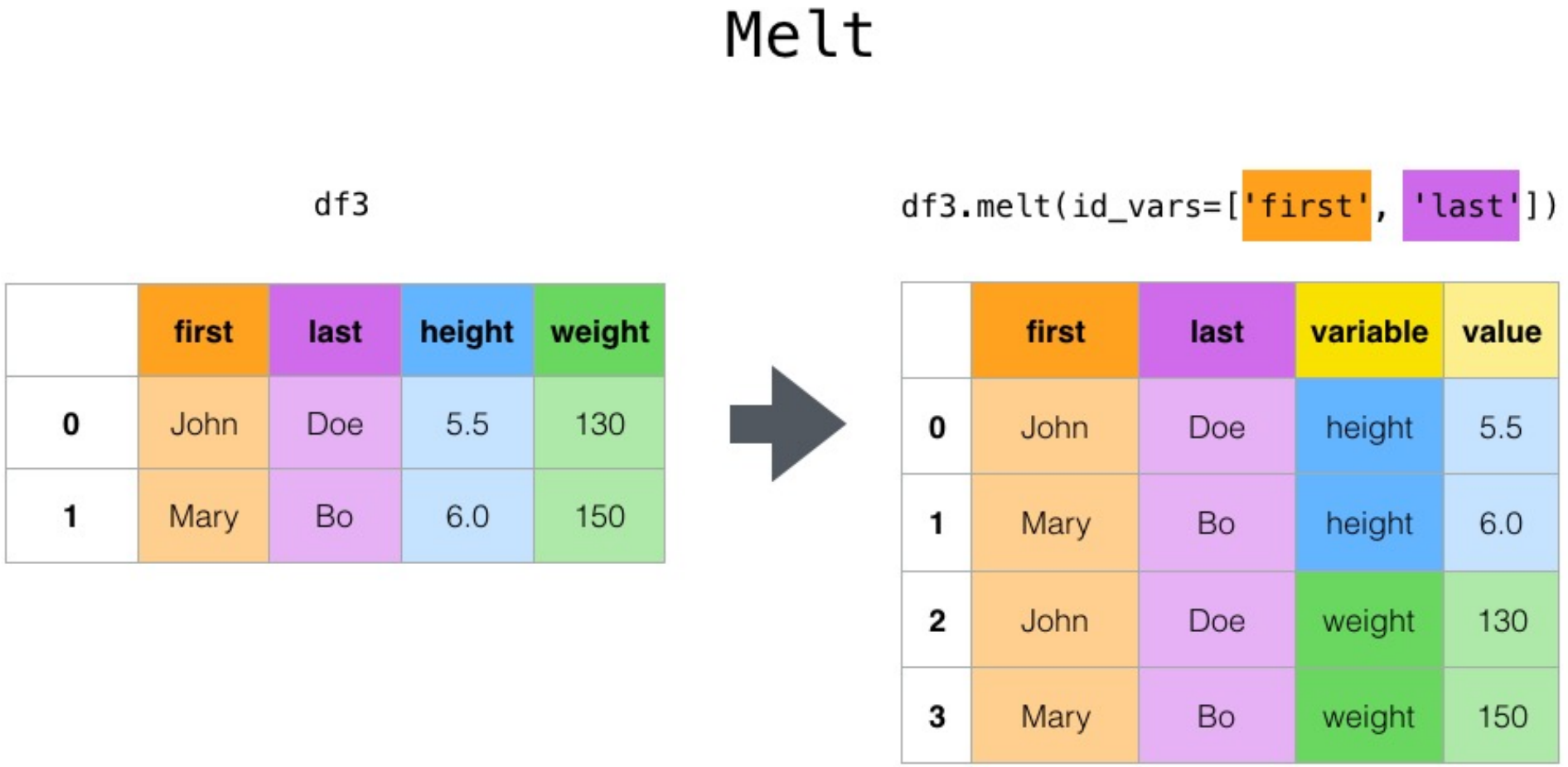
- Tidy data to wide data: `df.pivot` or `df.unstack` (with `df.set_index()`)
  - **Pivot/unmelt strategy**: Use `.pivot()` method on melted DataFrame and specify indices and columns
    - Shouldn't need to worry about aggfunc → all observations should be preserved (just in a different shape)
  - **Unstack strategy**: `.set_index()` on id *and* value columns, `.unstack()` value columns

pandas data reshaping: `stack`



http://pandas.pydata.org/pandas-docs/stable/user_guide/reshaping.html

pandas data reshaping: `melt`

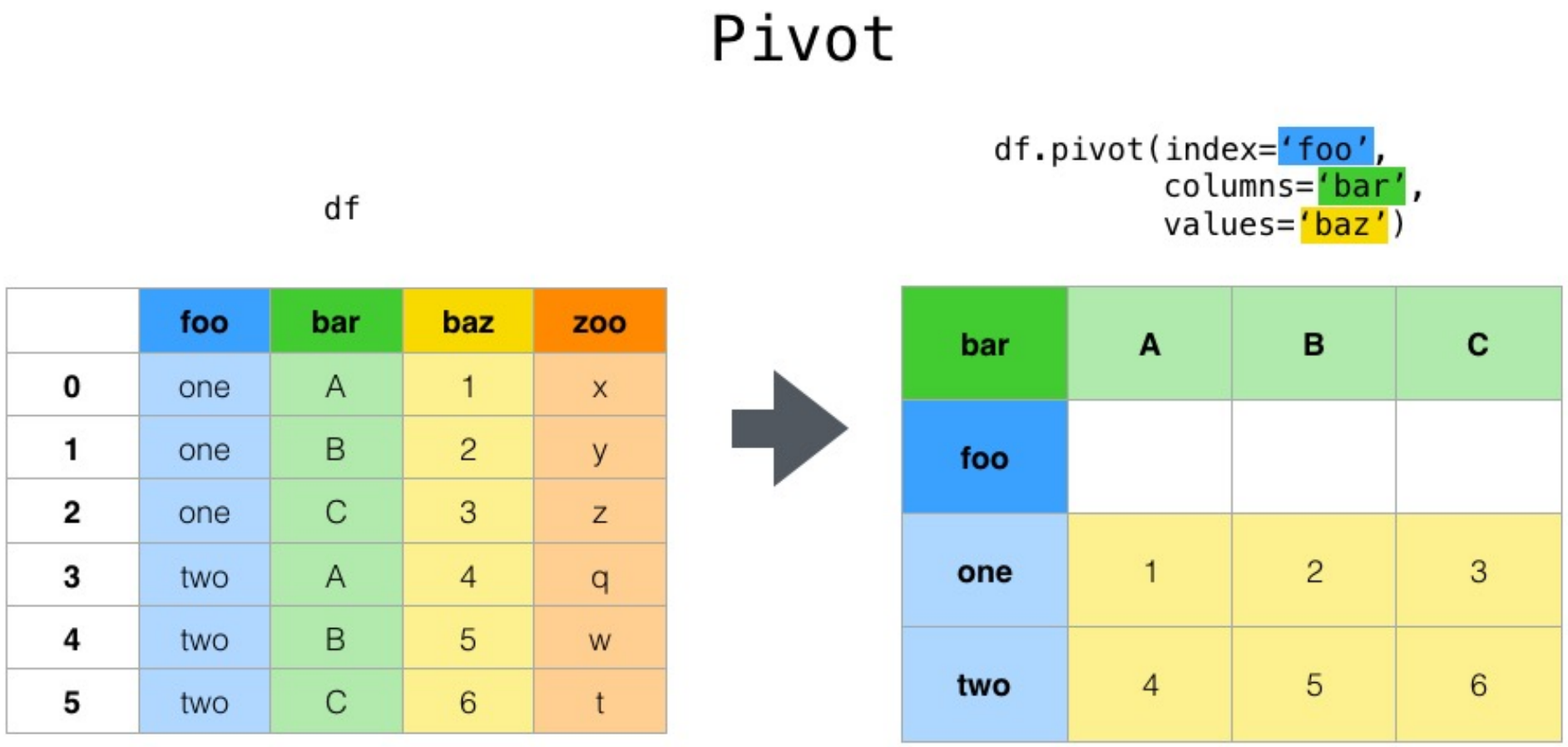


	first	last	height	weight
0	John	Doe	5.5	130
1	Mary	Bo	6.0	150

	first	last	variable	value
0	John	Doe	height	5.5
1	Mary	Bo	height	6.0
2	John	Doe	weight	130
3	Mary	Bo	weight	150

http://pandas.pydata.org/pandas-docs/stable/user_guide/reshaping.html

pandas data reshaping: `pivot`

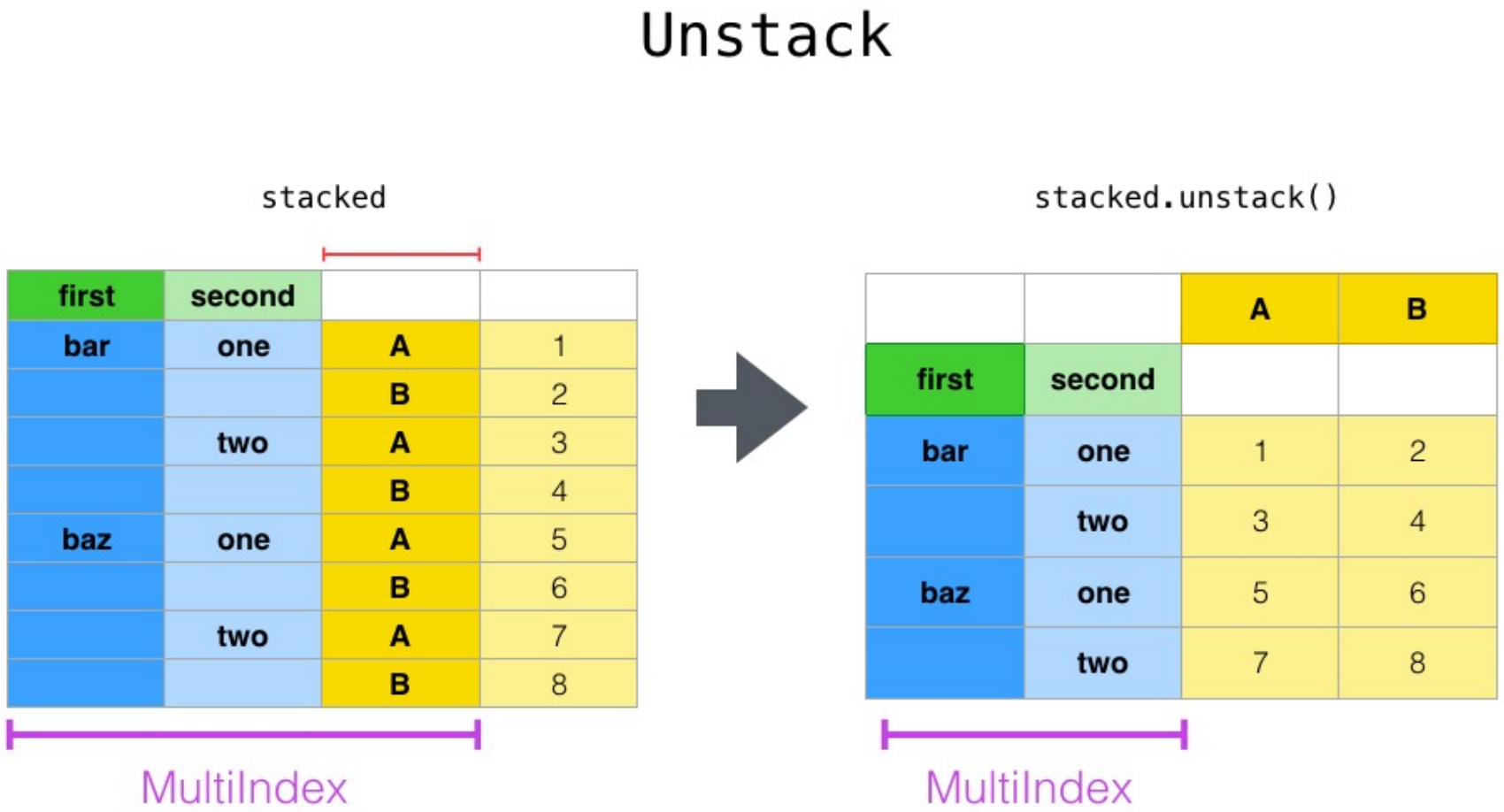


	foo	bar	baz	zoo
0	one	A	1	x
1	one	B	2	y
2	one	C	3	z
3	two	A	4	q
4	two	B	5	w
5	two	C	6	t

bar	A	B	C
foo			
one	1	2	3
two	4	5	6

http://pandas.pydata.org/pandas-docs/stable/user_guide/reshaping.html

pandas data reshaping: `unstack`



http://pandas.pydata.org/pandas-docs/stable/user_guide/reshaping.html

Notebook time!

- Download “Week 04 – Lecture.ipynb”, “flights.csv”, “transstats_passengers.xlsx”

Next class

Next Class

- Review concepts and exercises from last class
  - Complete “Thursday Questions” form! https://forms.gle/r3gsfjeY4pFM8t8M7 (ungraded/optional)
- Critique a data narrative or visualization
  - Post a link and a few sentences about a data visualization on Canvas discussion (ungraded/optional)
- Time to work on Weekly Assignment
  - Review results of Moving Deadline poll on Thursday https://forms.gle/rMtyjrrXpkEUeuk37
  - Weekly Assignment due on Friday* by submitting HTML notebook to Canvas before midnight
- Weekly quiz at the end of class (12:00–12:30)