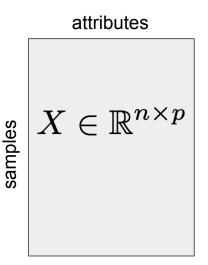
Data Transformations Part 2: Tidy Data

IFT6758 Fall 2019

More than just tricks

- Last time: Data types, missingness, joining and reorganizing
- A deeper principle: *Tidying*

Structuring datasets to facilitate analysis

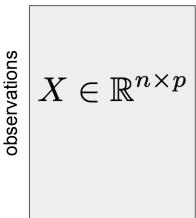


What are tidy data?

- 1. Variables are in columns.
- 2. Observations are in rows.
- 3. Different observational units types \rightarrow Different tables

(just renaming, they mean the same thing)

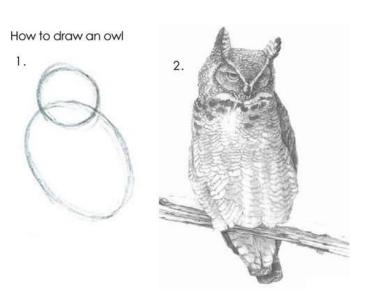
variables



What are tidy data?

- 1. Variables are in columns.
- 2. Observations are in rows.
- 3. Different observational units types \rightarrow Different tables

(example in pew.ipynb)



(just renaming, they mean the same thing)

observations

variables

 $X \in \mathbb{R}^{n \times p}$

Subtlety: Observation Amibuity

What are considered variables vs. observations may vary throughout your analysis.

Rules of thumb

- Functional relationships are easiest to see through columns
- Group comparisons are easiest to see through rows
- An observation is the smallest unit you'd like to draw conclusions or make predictions about

 $X \in \mathbb{R}^{n \times p}$

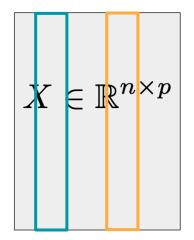
Subtlety: Observation Amibuity

What are considered variables vs. observations may vary throughout your analysis.

Rules of thumb

- Functional relationships are easiest to see through columns
- Group comparisons are easiest to see through rows
- An observation is the smallest unit you'd like to draw conclusions or make predictions about

10 * x[4] + 2 * x[6]

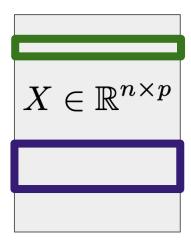


Subtlety: Observation Ambiguity

What are considered variables vs. observations may vary throughout your analysis.

Rules of thumb

- Functional relationships are easiest to see through columns
- Group comparisons are easiest to see through rows
- An observation is the smallest unit you'd like to draw conclusions or make predictions about



Example: Billboard Rankings

What makes more sense as an observation?

Option A: Each song Option B: Each song's timepoint Sorry, indie fans. But maybe check out... https://www.kaggle.com/nolanbconaway /pitchfork-data

(go through example in billboard.ipynb)

We'll see how both are relevant, and how to reconcile them.

		The second s								
jenre	date.entered	date.peaked	x1st.week	x2nd.week	x3rd_week		x67th.week	x68th.week	x69th.week	x70th.week
Rock	2000-09-23	2000-11-18	78	63.0	49.0		NaN	NaN	NaN	NaN
Rock	2000-02-12	2000-04-08	15	8.0	6.0		NaN	NaN	NaN	NaN
Rock	1999-10-23	2000-01-29	71	48.0	43.0		NaN	NaN	NaN	NaN
Rock	2000-08-12	2000-09-16	41	23.0	18.0		NaN	NaN	NaN	NaN
Rock	2000-08-05	2000-10-14	57	47.0	45.0		NaN	NaN	NaN	NaN
R&B	2000-08-05	2000-08-05	98	NaN	NaN		NaN	NaN	NaN	NaN
Rap	2000-02-12	2000-02-12	99	99.0	99.0		NaN	NaN	NaN	NaN
Rock	2000-09-02	2000-09-02	99	99.0	NaN		NaN	NaN	NaN	NaN
Rap	2000-07-01	2000-07-01	99	99.0	NaN		NaN	NaN	NaN	NaN
R&B	2000-10-28	2000-10-28	99	NaN	NaN		NaN	NaN	NaN	NaN
	Rock Rock Rock Rock Rock R&B Rap Rock Rap	Rock 2000-09-23 Rock 2000-02-12 Rock 1999-10-23 Rock 2000-08-12 Rock 2000-08-05 R&B 2000-08-05 Rap 2000-02-12 Rock 2000-08-05 Rap 2000-02-12 Rock 2000-09-02 Rap 2000-07-01	Rock2000-09-232000-11-18Rock2000-02-122000-04-08Rock1999-10-232000-01-29Rock2000-08-122000-09-16Rock2000-08-052000-10-14R&B2000-08-052000-08-05Rap2000-02-122000-02-12Rock2000-09-022000-02-12Rock2000-09-022000-09-02Rap2000-07-012000-07-01	Rock 2000-09-23 2000-11-18 78 Rock 2000-02-12 2000-04-08 15 Rock 1999-10-23 2000-01-29 71 Rock 2000-08-12 2000-09-16 41 Rock 2000-08-05 2000-10-14 57 R&B 2000-08-05 2000-08-05 98 Rap 2000-02-12 2000-02-12 99 Rock 2000-09-02 2000-09-02 99 Rap 2000-07-01 2000-07-01 99	Rock 2000-09-23 2000-11-18 78 63.0 Rock 2000-02-12 2000-04-08 15 8.0 Rock 1999-10-23 2000-01-29 71 48.0 Rock 2000-08-12 2000-09-16 41 23.0 Rock 2000-08-05 2000-10-14 57 47.0 R&B 2000-08-05 98 NaN Rap 2000-02-12 2000-02-12 99 99.0 Rock 2000-09-02 2000-09-02 99 99.0 Rap 2000-07-01 2000-07-01 99 99.0	Rock 2000-09-23 2000-11-18 78 63.0 49.0 Rock 2000-02-12 2000-04-08 15 8.0 6.0 Rock 1999-10-23 2000-01-29 71 48.0 43.0 Rock 2000-08-12 2000-09-16 41 23.0 18.0 Rock 2000-08-05 2000-10-14 57 47.0 45.0 R&B 2000-08-05 2000-08-05 98 NaN NaN Rap 2000-02-12 2000-02-12 99 99.0 99.0 Rock 2000-09-02 2000-09-02 99 99.0 NaN Rap 2000-07-01 2000-07-01 99 99.0 NaN	Rock 2000-09-23 2000-11-18 78 63.0 49.0 Rock 2000-02-12 2000-04-08 15 8.0 6.0 Rock 1999-10-23 2000-01-29 71 48.0 43.0 Rock 2000-08-12 2000-09-16 41 23.0 18.0 Rock 2000-08-05 2000-10-14 57 47.0 45.0 R&B 2000-08-05 2000-08-05 98 NaN NaN Rap 2000-02-12 2000-02-12 99 99.0 99.0 Rock 2000-09-02 2000-09-02 99 99.0 NaN Rap 2000-07-01 2000-07-01 99 99.0 NaN	Rock 2000-09-23 2000-11-18 78 63.0 49.0 NaN Rock 2000-02-12 2000-04-08 15 8.0 6.0 NaN Rock 1999-10-23 2000-01-29 71 48.0 43.0 NaN Rock 2000-08-12 2000-09-16 41 23.0 18.0 NaN Rock 2000-08-05 2000-10-14 57 47.0 45.0 NaN R&B 2000-08-05 2000-08-05 98 NaN NaN NaN Rap 2000-02-12 2000-02-12 99 99.0 99.0 NaN Rock 2000-09-02 2000-09-02 99 99.0 NaN NaN Rap 2000-07-01 2000-07-01 99 99.0 NaN NaN	Rock 2000-09-23 2000-11-18 78 63.0 49.0 NaN NaN NaN Rock 2000-02-12 2000-04-08 15 8.0 6.0 NaN NaN NaN Rock 1999-10-23 2000-01-29 71 48.0 43.0 NaN NaN NaN Rock 2000-08-12 2000-09-16 41 23.0 18.0 NaN NaN NaN Rock 2000-08-05 2000-10-14 57 47.0 45.0 NaN NaN NaN Rock 2000-08-05 2000-08-05 98 NaN NaN NaN NaN Rap 2000-02-12 2000-02-12 99 99.0 99.0 NaN NaN NaN Rock 2000-09-02 2000-09-02 99 99.0 NaN NaN NaN Rock 2000-09-02 2000-09-02 99 99.0 NaN NaN NaN Rock 2000-07-01 2000-07-01 99 99.0 NaN NaN NaN	Rock 2000-09-23 2000-11-18 78 63.0 49.0 NaN NaN NaN NaN Rock 2000-02-12 2000-04-08 15 8.0 6.0 NaN NaN NaN NaN Rock 1999-10-23 2000-01-29 71 48.0 43.0 NaN NaN NaN NaN Rock 2000-08-12 2000-09-16 41 23.0 18.0 NaN NaN NaN NaN Rock 2000-08-05 2000-10-14 57 47.0 45.0 NaN NaN NaN NaN Rock 2000-08-05 2000-08-05 98 NaN NaN NaN NaN NaN Rap 2000-02-12 2000-02-12 99 99.0 99.0 NaN NaN NaN NaN Rap 2000-09-02 2000-09-02 99 99.0 NaN NaN NaN NaN Rap 2000-07-01 2000-07-01 99 99.0 NaN NaN NaN NaN