Data Transformations Part 2:

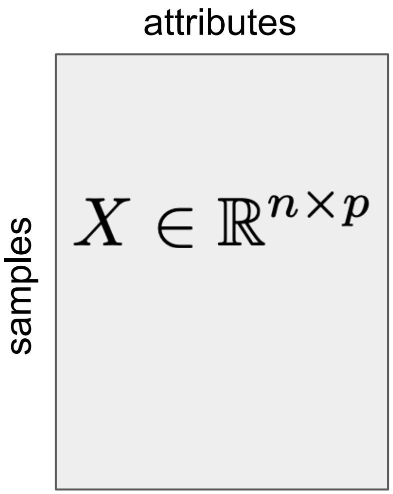
Tidy Data

IFT6758
Fall 2019

More than just tricks

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

Structuring datasets to facilitate analysis



attributes

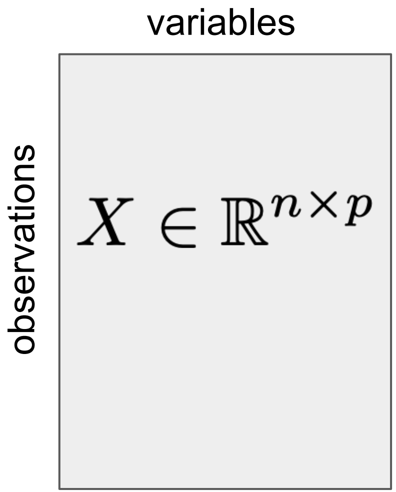
samples

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

What are tidy data?

(just renaming, they mean the same thing)

1. Variables are in columns.
2. Observations are in rows.
3. Different observational units types → Different tables



variables

observations

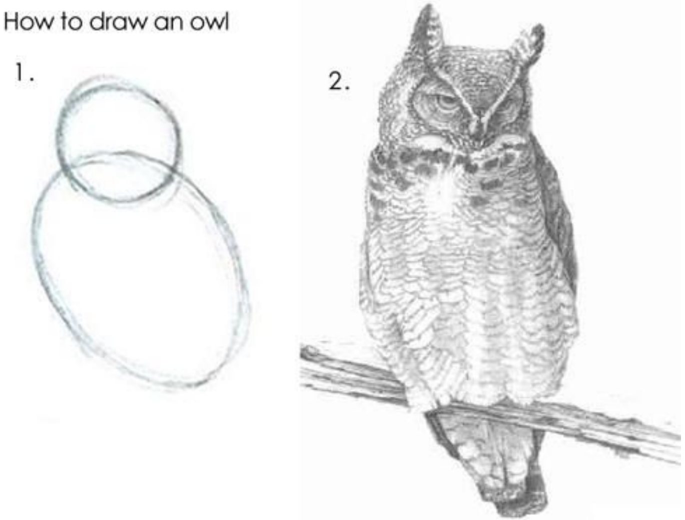
$X \in \mathbb{R}^{n \times p}$

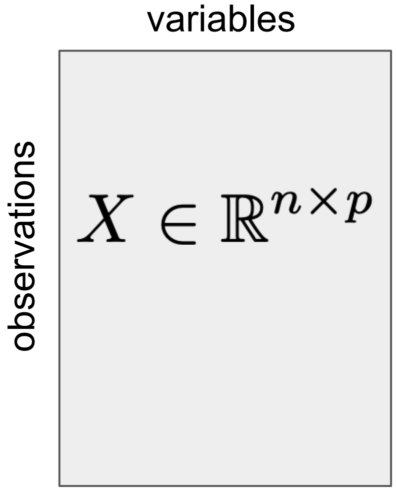
What are tidy data?

(just renaming, they mean the same thing)

1. Variables are in columns.
2. Observations are in rows.
3. Different observational units types → Different tables

(example in pew.ipynb)





variables

observations

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

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

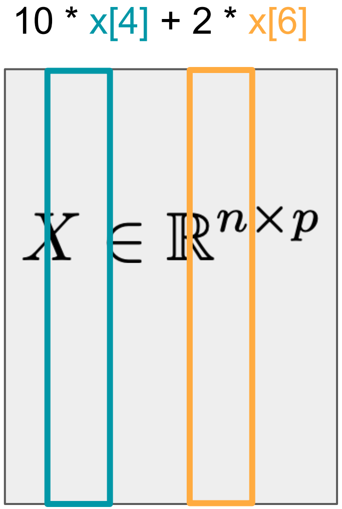
$$X \in \mathbb{R}^{n \times p}$$

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

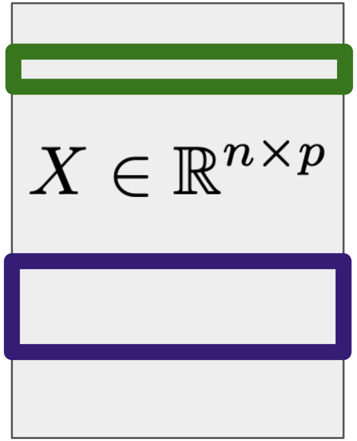


Subtlety: Observation Ambiguity

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



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

Example: Billboard Rankings

Sorry, indie fans. But maybe check out...
https://www.kaggle.com/nolanbconaway/pitchfork-data

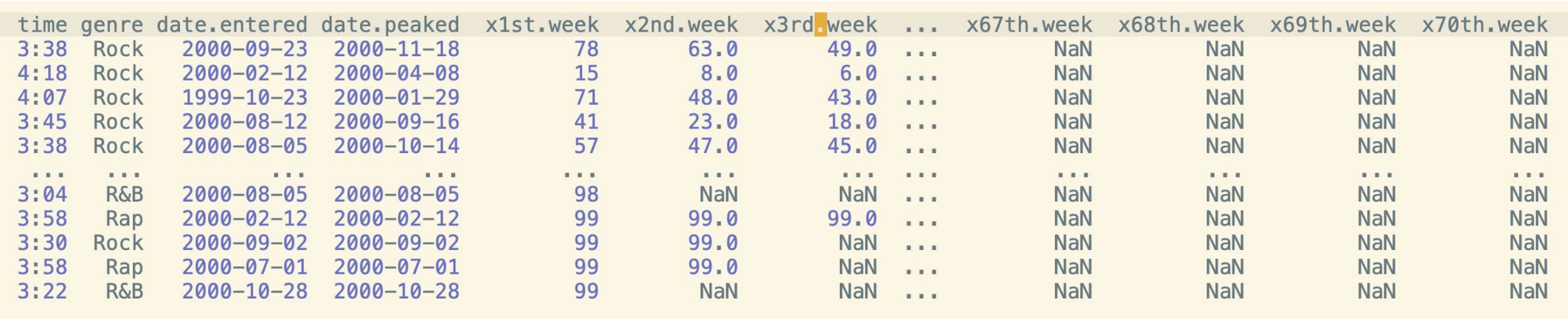
What makes more sense as an observation?

Option A: Each song

Option B: Each song’s timepoint

(go through example in billboard.ipynb)

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



```
time genre date.entered date.peaked  x1st.week  x2nd.week  x3rd.week  ...  x67th.week  x68th.week  x69th.week  x70th.week
3:38  Rock   2000-09-23  2000-11-18         78       63.0       49.0  ...         NaN         NaN         NaN         NaN
4:18  Rock   2000-02-12  2000-04-08         15        8.0        6.0  ...         NaN         NaN         NaN         NaN
4:07  Rock   1999-10-23  2000-01-29         71       48.0       43.0  ...         NaN         NaN         NaN         NaN
3:45  Rock   2000-08-12  2000-09-16         41       23.0       18.0  ...         NaN         NaN         NaN         NaN
3:38  Rock   2000-08-05  2000-10-14         57       47.0       45.0  ...         NaN         NaN         NaN         NaN
 ...   ...          ...         ...        ...        ...        ...  ...         ...         ...         ...         ...
3:04   R&B   2000-08-05  2000-08-05         98        NaN        NaN  ...         NaN         NaN         NaN         NaN
3:58   Rap   2000-02-12  2000-02-12         99       99.0       99.0  ...         NaN         NaN         NaN         NaN
3:30  Rock   2000-09-02  2000-09-02         99       99.0        NaN  ...         NaN         NaN         NaN         NaN
3:58   Rap   2000-07-01  2000-07-01         99       99.0        NaN  ...         NaN         NaN         NaN         NaN
3:22   R&B   2000-10-28  2000-10-28         99        NaN        NaN  ...         NaN         NaN         NaN         NaN
```