CLEANING DATA IN R

Introduction to tidy data

Principles of tidy data

name	age	eye_color	height
Jake	34	Other	6'1"
Alice	55	Blue	5'9"
Tim	76	Brown	5'7"
Denise	19	Other	5'1"

Observation

Variable or Attribute

- Observations as rows
- Variables as columns
- One type of observational unit per table

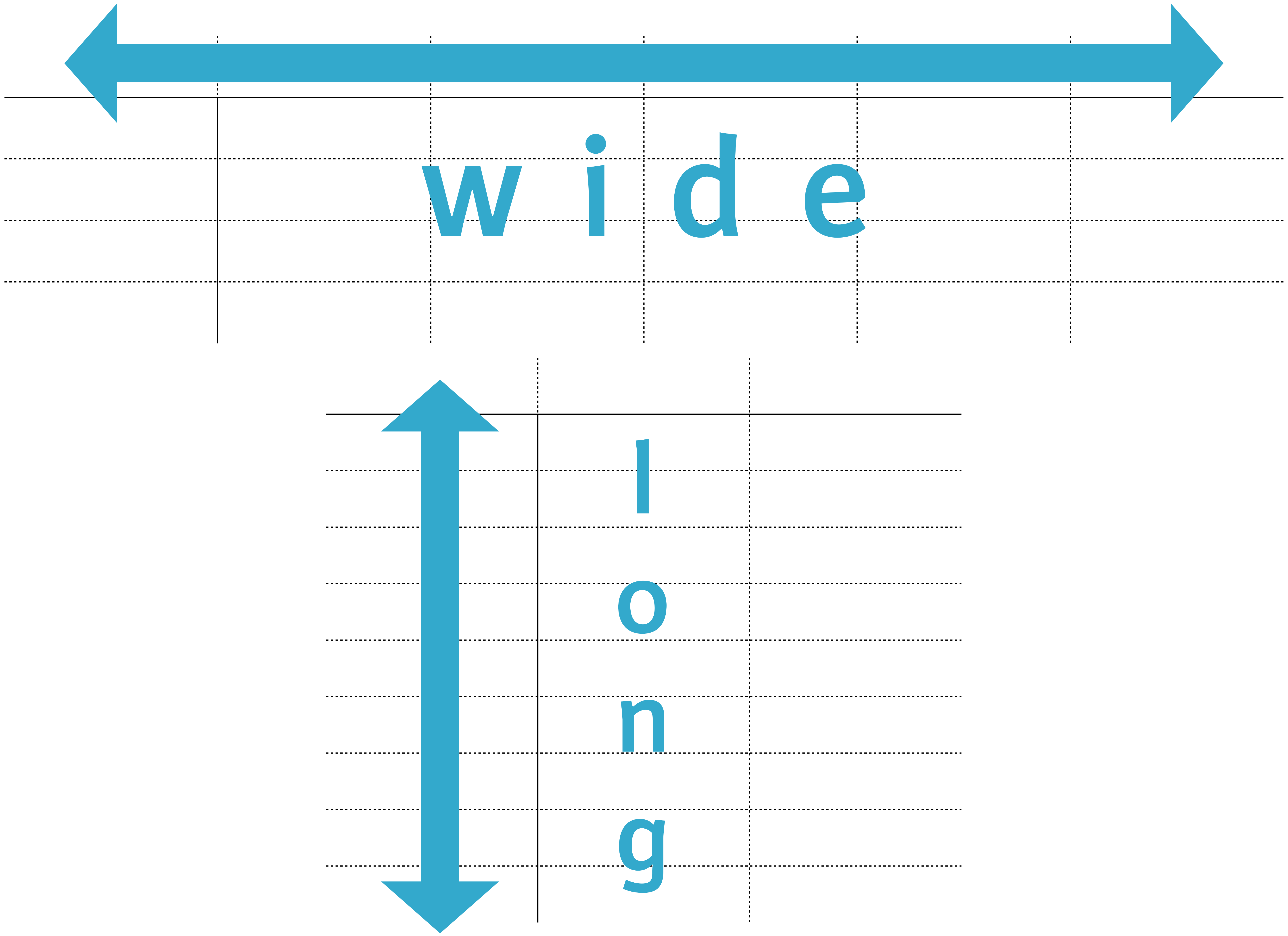
A dirty data diagnosis

name	age	brown	blue	other	height
Jake	34	0	0	1	6'1"
Alice	55	0	1	0	5'9"
Tim	76	1	0	0	5'7"
Denise	19	0	0	1	5'1"

?

Column headers are values, not variable names

Wide vs. long datasets



Let's practice!

CLEANING DATA IN R

Introduction to tidyr

Overview of tidyr

- R package by Hadley Wickham
- Apply the principles of tidy data
- Small set of simple functions

Gather columns into key-value pairs

```
# Look at wide_df
> wide_df
  col A B C
1   X 1 2 3
2   Y 4 5 6

# Gather the columns of wide_df
> gather(wide_df, my_key, my_val, -col)
  col my_key my_val
1   X      A      1
2   Y      A      4
3   X      B      2
4   Y      B      5
5   X      C      3
6   Y      C      6
```

`gather(data, key, value, …)`

`data`: a data frame

`key`: bare name of new key column

`value`: bare name of new value column

…: bare names of columns to gather (or not)

Spread key-value pairs into columns

```
# Look at long_df
> long_df
  col my_key my_val
1   X      A      1
2   Y      A      4
3   X      B      2
4   Y      B      5
5   X      C      3
6   Y      C      6

# Spread the key-value pairs of long_df
> spread(long_df, my_key, my_val)
  col A B C
1   X 1 2 3
2   Y 4 5 6
```

`spread(data, key, value)`

`data`: a data frame

`key`: bare name of column containing keys

`value`: bare name of column containing values

Let's practice!

CLEANING DATA IN R

Introduction to tidyr

Separate columns

```
# View the treatments data
> treatments
  patient treatment year_mo response
1       X         A 2010-10        1
2       Y         A 2010-10        4
3       X         B 2012-08        2
4       Y         B 2012-08        5
5       X         C 2014-12        3
6       Y         C 2014-12        6

# Separate year_mo into two columns
> separate(treatments, year_mo, c("year", "month"))
  patient treatment year month response
1       X         A 2010    10        1
2       Y         A 2010    10        4
3       X         B 2012    08        2
4       Y         B 2012    08        5
5       X         C 2014    12        3
6       Y         C 2014    12        6
```

`separate(data, col, into)`

`data`: a data frame

sep = "-"

`col`: bare name of column to separate

`into`: character vector of new column names

Unite columns

```
# View treatments data
> treatments
  patient treatment year month response
1       X         A 2010    10        1
2       Y         A 2010    10        4
3       X         B 2012    08        2
4       Y         B 2012    08        5
5       X         C 2014    12        3
6       Y         C 2014    12        6

# Unite year and month to form year_mo column
> unite(treatments, year_mo, year, month)
  patient treatment year_mo response
1       X         A 2010_10        1
2       Y         A 2010_10        4
3       X         B 2012_08        2
4       Y         B 2012_08        5
5       X         C 2014_12        3
6       Y         C 2014_12        6
```

`unite(data, col, …)`

`data`: a data frame **sep = "-"**

`col`: bare name of new column

…: bare names of columns to unite

Summary of key tidyr functions

- `gather()` - Gather columns into key-value pairs
- `spread()` - Spread key-value pairs into columns
- `separate()` - Separate one column into multiple
- `unite()` - Unite multiple columns into one

Let's practice!

CLEANING DATA IN R

Common symptoms of messy data

Column headers are values, not variable names

name	age	brown	blue	other	height
Jake	34	0	0	1	6'1"
Alice	55	0	1	0	5'9"
Tim	76	1	0	0	5'7"
Denise	19	0	0	1	5'1"

name	age	eye_color	height
Jake	34	Other	6'1"
Alice	55	Blue	5'9"
Tim	76	Brown	5'7"
Denise	19	Other	5'1"

Variables are stored in both rows and columns

name	measurement	value
Jake	n_dogs	1
Jake	n_cats	0
Jake	n_birds	1
Alice	n_dogs	1
Alice	n_cats	2
Alice	n_birds	0

name	n_dogs	n_cats	n_birds
Jake	1	0	1
Alice	1	2	0

Multiple variables are stored in one column

name	sex_age	eye_color	height
Jake	M.34	Other	6'1"
Alice	F.55	Blue	5'9"
Tim	M.76	Brown	5'7"
Denise	F.19	Other	5'1"

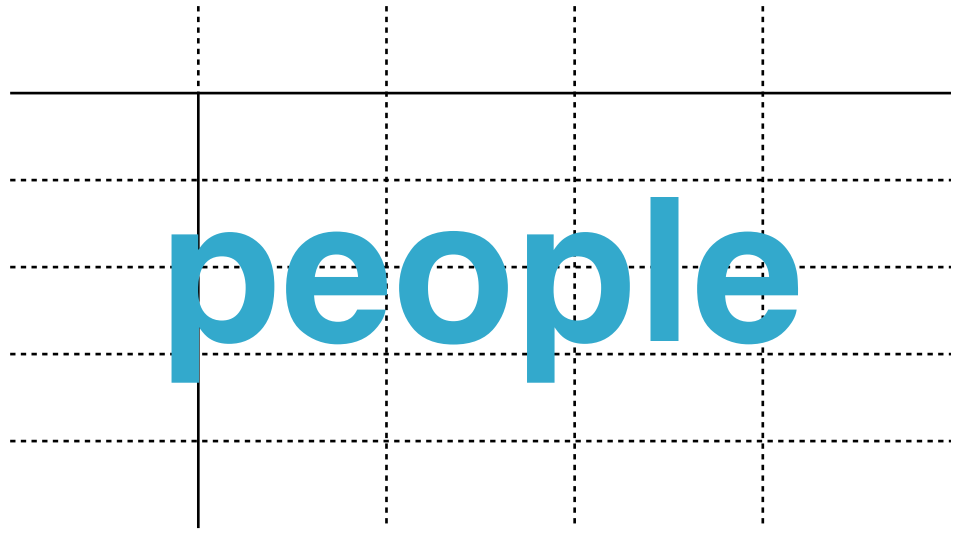
name	sex	age	eye_color	height
Jake	M	34	Other	6'1"
Alice	F	55	Blue	5'9"
Tim	M	76	Brown	5'7"
Denise	F	19	Other	5'1"

Other common symptoms

- A single observational unit is stored in multiple tables
- Multiple types of observational units are stored in the same table

name	age	height	pet_name	pet_type	pet_height
Jake	34	6'1"	Larry	Dog	25"
Jake	34	6'1"	Chirp	Bird	3"
Alice	55	5'9"	Wally	Dog	30"
Alice	55	5'9"	Sugar	Cat	10"
Alice	55	5'9"	Spice	Cat	12"

Alice's name, age, and height are duplicated 3x



CLEANING DATA IN R

Let's practice!