🧹 Tidy Data

Semantics

- A dataset is a collection of **values**.
- Every value belongs to a variable and an observation.
- A **variable** contains all values that measure the same underlying attribute.
- An **observation** contains all values measured on the same unit.

What is Tidy Data?

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

Source: Hadley Wickham, Tidy Data

Common Mistakes

1. Column headers are **values**, not variable names.
2. Multiple variables are **stored in one column**.
3. A combination of the first two mistakes.
4. Data on the wrong aggregation level.

Key Question: What are the necessary inputs for my ML algorithm to predict one value?

Error Type 1: Headers as Values

Messy

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

Mistake: My ML algorithm wants to predict the result based on the treatment type. In each row, there are two instances of my observation.

Error Type 1: Headers as Values

Tidy

name	trt	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

Correction: Now one observation per row.

Your task

Please clean up the data on sheets `Bad_Example1` `Bad_Example2` in the file `TidyData.xlsx`

Time: 10 Minutes

Hints: Use either melt or stack

Error Type 2: Multiple Variables in One Column

Messy Format

machine	date	key	value
M	1/1/22	tmin	15,00
M	1/1/22	tmax	17,00
M	1/2/22	tmin	15,20
M	1/2/22	tmax	17,40
M	1/3/22	tmin	15,90
M	1/3/22	tmax	18,30

Mistake: My ML algorithm wants to predict the result based on minimum temperature and maximum temperature. Each observation is split across two columns.

Error Type 2: Multiple Variables in One Column

Tidy Format

machine	date	tmin	tmax
M	1/1/22	15	17,00
M	1/2/22	15,20	17,40
M	1/3/22	15,90	18,30

Correction: Now one observation per row.

Your task

Please clean up the data on the first sheet of the file `TidyData_ErrorType_Two.xlsx`

Time: 10 Minutes

Hints: Use pivot

Error Type 4: Wrong aggregate level

Assume you have data on transaction level:

Messy format

Store	Date	Transaction ID	Amount
1	1/3/22	kdjvi	787,00
1	1/3/22	kjdfj	1887,40
2	1/3/22	qeoku	1148,30
2	1/3/22	jhkjg	87,30
2	1/3/22	phljh	8398,00
2	1/3/22	iohkl	118,90
2	1/3/22	drfhg	81,35
2	1/3/22	oekjj	1148,30

Mistake: You want predictions on sales per store per month.

Error Type 4: Wrong aggregate level

Tidy format

Store	Month	Amount
1	3/22	15787,00
2	3/22	82971,00

Correction: Performed Aggregates per month.

Your task

Please clean up the data

`transaction_data.csv`

Your prediction unit is item code and country

Time: 10 Minutes

Hints: Use groupby