

Week 3: Tidy Data

September 9, 2026

TIDY DATA = LONG FORMAT

Tidy: every **variable** has its own **column**, every **observation** has its own **row**. In practice that means **long** format — one row per observation, with a column naming *what* is being measured, not one column per category.

	“Wide”	“Long” (tidy)
Shape	One column per department	One row per department per year
Columns	year, DHS, DOC, DOD, ...	year, department, rd_budget_mil
Reads naturally as	A spreadsheet for humans	Data dplyr/ggplot can group on

The tell: if a column name is a *value* (a department, a gender, a film) rather than a description of what’s in the column, the table is wide — e.g. TB cases with one column per country, or LOTR word counts with Female/Male columns. Ask “does this column name describe the values under it?” — yes → long/tidy, no → wide.

RESHAPING: PIVOT_LONGER() / PIVOT_WIDER()

Function	Direction	Key arguments
<code>pivot_longer()</code>	wide → long	<code>cols</code> to stack, <code>names_to/</code> <code>values_to</code> to name the new columns
<code>pivot_wider()</code>	long → wide	<code>names_from/values_from</code> to supply the new column names and values

```
# wide -> long
fed_spend_long <- fed_spend_wide %>%
  pivot_longer(
    names_to = "department",
    values_to = "rd_budget_mil",
    cols = -year # every column except
  year
  )

# long -> wide
fed_spend_wide <- fed_spend_long %>%
  pivot_wider(
    names_from = department,
    values_from = rd_budget_mil
  )
```

`cols` takes either the columns *to include* (`DOD:Other`) or, often shorter, the ones *to exclude* (`-year`) — pick whichever names fewer columns.

WRANGLING: RESHAPE FIRST, THEN SUMMARIZE

Computing “total spending per year” on wide data means writing out every department by hand (`DHS + DOC + DOD + ...`) — it breaks the moment a column is added or renamed. On the long `fed_spend_long` from above, it’s one `group_by()` + `summarise()`:

```
fed_spend_long %>%
  group_by(year) %>%
  summarise(total = sum(rd_budget_mil))
```

If you’re typing out every column name in a formula, you’re in the wrong shape. dplyr and ggplot both expect long data — reshape before you wrangle or plot, not after.

VISUALIZING TIDY DATA

`geom_col(aes(x = rd_budget_mil, y = department))` fails on wide data — there is no `department` column to map, only `DHS`, `DOC`, `DOD`, ... as separate columns. `pivot_longer()` first turns each

department into a *value* in one column, which is what `aes()` needs to map a category to an axis, fill, or color.

DATA PROVENANCE & CURATION

Three questions to ask of any dataset before you trust it:

Angle	Ask
Validity	Trustworthy? Where from, and how managed over time?
Comprehension	Accurate, explainable, and the right data for the question?
Reproducibility	Could someone fully replicate your results from your raw data and code?

Document like a museum curator. Record: file name, download date, original file name (if renamed), URL, source of the *original* data (if different), how it was collected, and a data dictionary (one row per column).

WRITING A RESEARCH QUESTION

Trait	Means
Clear	Understandable without extra explanation
Focused	Narrow enough to answer with the data and time available
Concise	Expressed in the fewest words possible
Complex	Not answerable with a simple yes/no — needs analysis
Arguable	Answers are open to debate, not an accepted fact

Bad: *Why are social networking sites harmful?* — unclear which sites, what “harm” means. **Better:** *How are online users experiencing privacy issues on sites such as Facebook and Twitter?* — names the sites, the harm, and who.

See past [example projects](#) for full worked examples.

ERRORS YOU MAY HIT THIS WEEK

Message / symptom	Usual cause
object 'department' not found in <code>aes()</code>	Data is still wide — <code>pivot_longer()</code> before mapping a category to <code>aes()</code>
“Values ... are not uniquely identified”	<code>pivot_wider()</code> found duplicate <code>names_from/</code> row-id combos — <code>summarise()</code> first

BEFORE NEXT CLASS

No reading reflection this week — focus on your **final project proposal**, due **Sept 20**.

- Form your team on the [team-formation sheet](#), if you haven’t.
- Sign up for a [1-on-1 proposal meeting](#) with Prof. Helveston.
- HW3: Cleaning Data** is posted, due in two weeks.

You should be able to:

- Say whether a table is long or wide, and reshape between them with `pivot_longer()/pivot_wider()`.
- Explain why dplyr and ggplot expect long data, and reshape before wrangling or plotting.
- Write a data README entry with source, date, and dictionary for a downloaded dataset.
- Write a research question that is clear, focused, concise, complex, and arguable.

If any of those are shaky, that’s what Slack and office hours are for.