

Week 1 • The foundation under data analysis

STAT163 — Data Manipulation Essentials

Oleh Omelchenko

2026-09-01

Agenda

- **01** What is this course about

- **02** How the course runs — schedule, grading, AI

- **03** What is a dataframe — one shape, many tools
 - What one column holds — four kinds of values
 - What one row means — grain and levels of detail

- **04** Before your practice — what to install

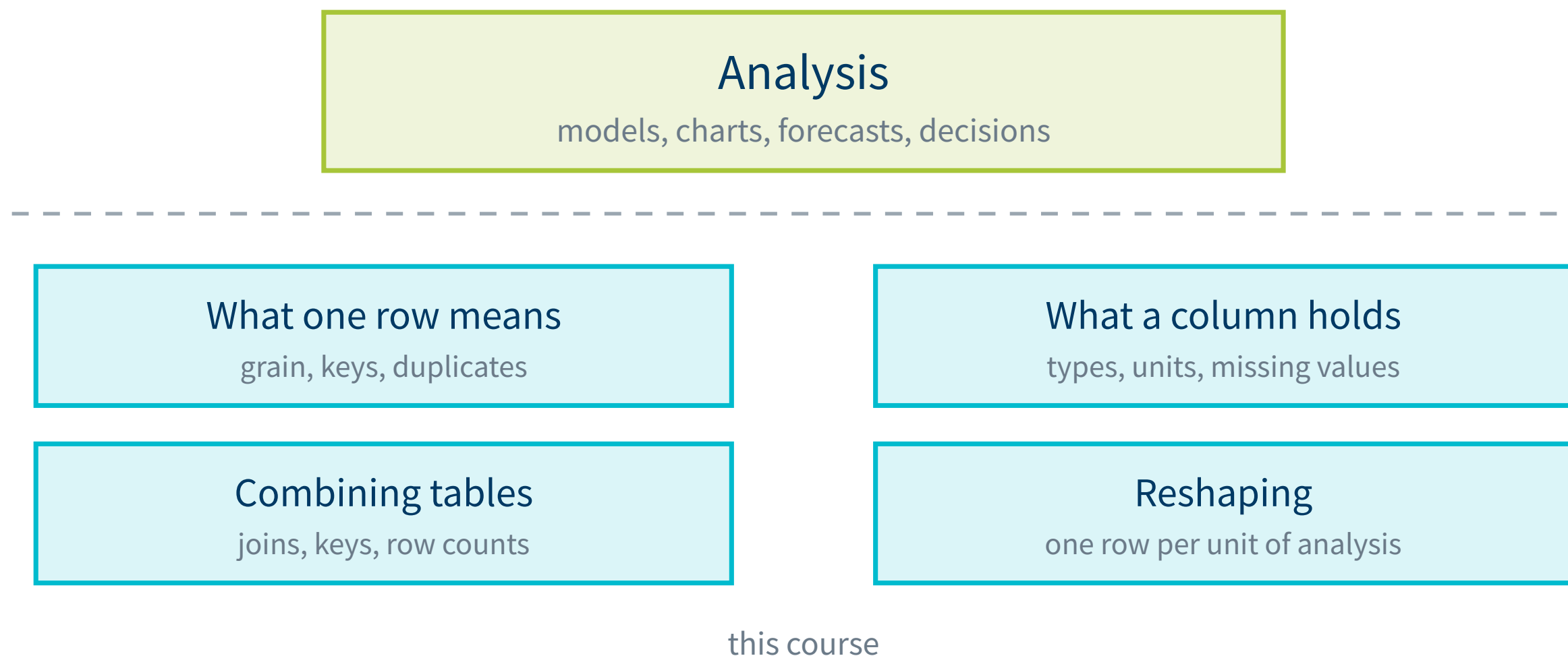
1 • What is this course about

What did you sign up for?

OPEN FLOOR

- What do you think this course teaches?
- The title says *data manipulation*. What gets manipulated, and into what?
- Where does that sit next to *data analysis* — before it, after it, part of it?

Analysis is the part you can see



Garbage in, garbage out: everything above the line is only as good as everything below it.

The goal

Look at any dataset and see its potential: what you can get out of it, what you cannot, and which techniques help you get from the table you have to the answer you need.

What you will be able to do

By November you can	Which means
Open an unfamiliar dataset and say what it holds	Shape, grain, types, what is missing, which values cannot be real
Reshape a table to answer a question	Filter, derive, group, pivot — and know which one you need
Combine tables without corrupting the result	Choose the key, and check the row count before and after
Handle missing values on purpose	Decide what to do with them and record why
Hand your work to someone else	They re-run it and get your number back

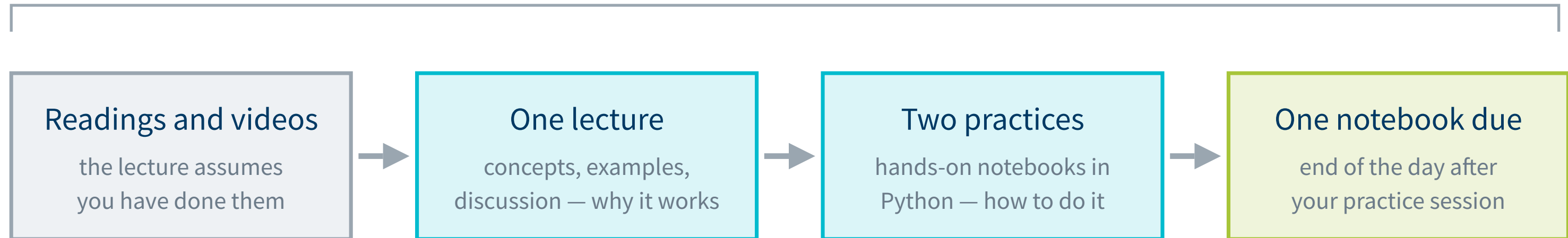
The tool is pandas

- The concepts are the course. Python and pandas are how we practise them.
- The same operations have the same names in spreadsheets, SQL and R.
- Learn an operation once and it transfers across tools.

2 • How the course runs

Eight weeks

every week



Grading

Component	Points
Practices (7 × 3)	21
Pre-reading quizzes (5 × 2)	10
Assignments (3)	24
Mid-term	20
Final	35
Total	110

AI in this course



PRACTICES AND ASSIGNMENTS

AI allowed, use it with disclosure



QUIZZES, MID-TERM, FINAL

No AI. Closed book

Use AI to learn, prove you learned without it.

Three rules for take-home work

Say what you used

One line in the notebook — the tool and the step.
"Used Claude to debug the groupby in Task 3.2."

You own what you submit

An AI error is your error. "The model wrote it" is not a defence.

Expect to explain it aloud

If you cannot say why your code does what it does, the exams will show it.

How I use AI

Slides and quiz questions

Drafted with Claude. I review, edit and iterate on every one.

Practices and assignments

Checked semi-automatically against an answer key. I review the output.

Exams

Read and graded by me, start to finish. No automation.

Assignments may come with a short talk

After any assignment, I (or TA) may pick a random submission and ask the submitter to talk me through their work for five to ten minutes: why this join, why this grain, why this number. The talk confirms or adjusts that assignment's mark.

3 · What is a dataframe

One shape, many names



The dataframe is the key unit we work with, in this course and beyond it. R, SQL and spreadsheets hold data shaped the same way, so everything you learn on it carries over.

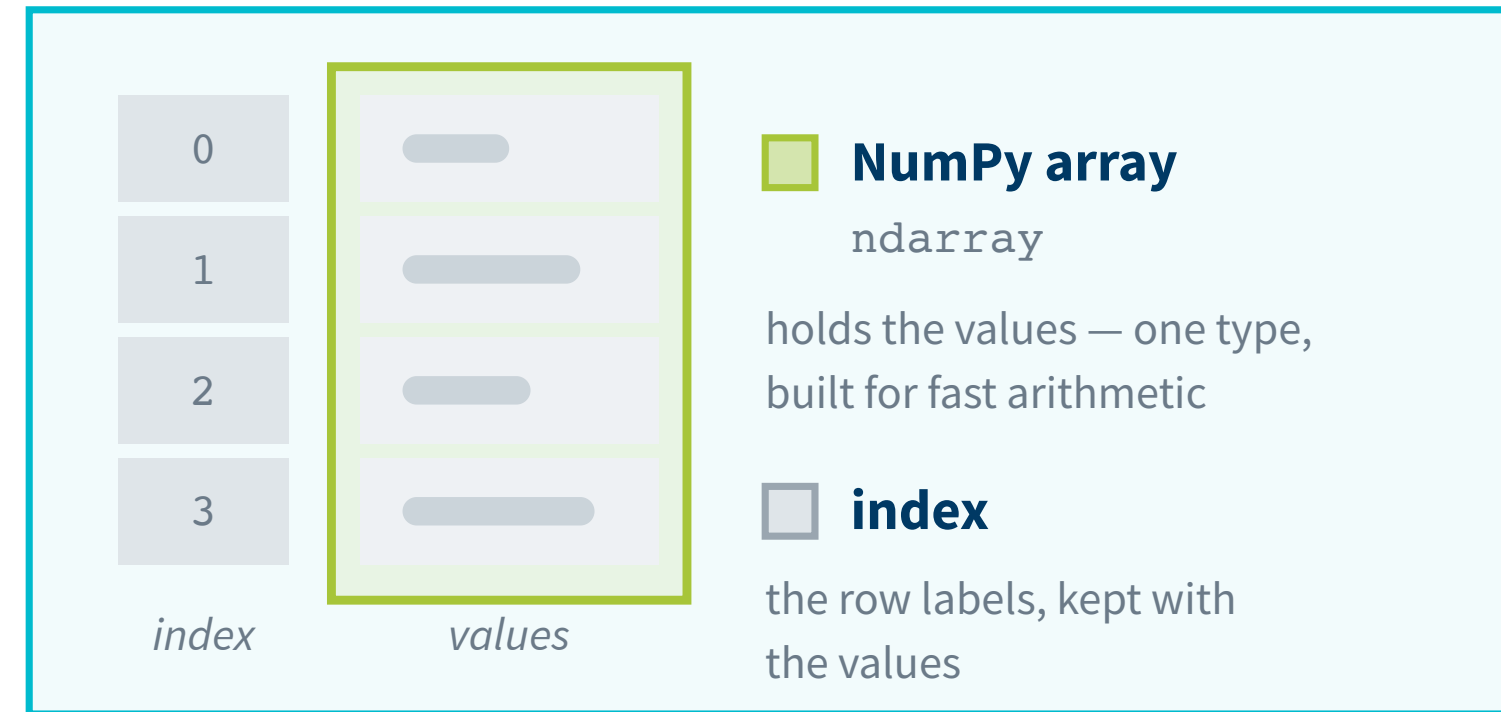
Whatever other data structures we'll encounter during this course, we will transform the data to be in tabular format.

A column is a Series

DataFrame



Series



A column pulled from a DataFrame is a Series. Inside a Series, the values usually sit in a NumPy array. A DataFrame is many Series sharing one index.

Columns and rows

■ A column

one property, recorded the same way for every row

■ A row

one observation — one delivery, one reading, one match

The rest of the lecture asks two questions of any table: what does one column hold, and what does one row mean?

3.1 · What one column holds

A delivery log

DELIVERIES — AS RECORDED

parcel	destination	weight	rating	delivered
P-1041	Lviv	2.4	5	2026-03-02
P-1042	Kyiv	1800 g	great	yesterday
P-1043	Odesa	heavy	★★★★★	2026-03-04
P-1044	Kyiv	0.8	4	last week
P-1045	Lviv	3.1	good	2026-03-05
P-1046	Dnipro	12	2	2026-03-06

TRY IT

- Total weight carried to Kyiv?
- Which destination has the best average rating?
- How many parcels arrived in the first week of March?

Every question fails on the same thing

Question	Why it fails
Total weight	2.4 , 1800 g and heavy are three different things in one column
Best average rating	Stars, words and numbers do not add
Parcels in week one	yesterday has no position on a calendar

As long as a column mixes kinds of value, almost nothing can be computed from it.

One kind of value per column

DELIVERIES — AFTER CLEANING

parcel	destination	weight_kg	rating	delivered
P-1041	Lviv	2.4	5	2026-03-02
P-1042	Kyiv	1.8	5	2026-03-03
P-1043	Odesa	11.0	4	2026-03-04
P-1044	Kyiv	0.8	4	2026-03-01
P-1045	Lviv	3.1	4	2026-03-05
P-1046	Dnipro	12.0	2	2026-03-06

heavy became **11.0** because someone decided what it meant. The table no longer shows that decision, which is why you write it down.

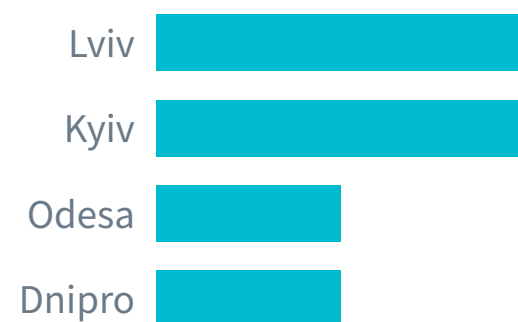
What you can do depends on the kind

parcel	destination	weight_kg	rating	delivered
P-1041	Lviv	2.4	5	2026-03-02
P-1042	Kyiv	1.8	4	2026-03-03
↓	↓	↓	↓	↓
a label, unique	a label, repeats	a measured amount	an ordered grade	a point in time
identifies the row — count, no arithmetic	group — average rating by city	sum, mean, difference	sort, median — a mean is arguable	order, range, "within the first week"

Four kinds

N Nominal

labels, no order

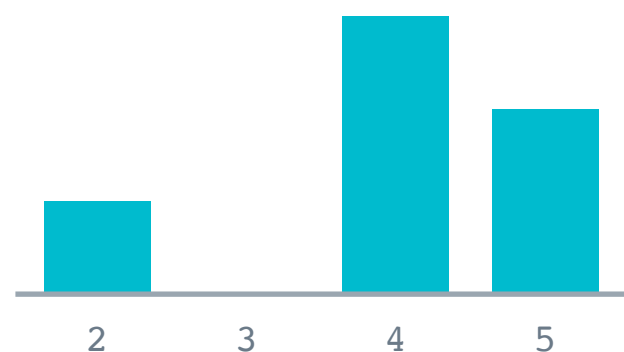


count per label —
any bar order works

parcel, destination

O Ordinal

ordered, gaps unknown



the order is fixed —
the gaps are unknown

rating, size S/M/L

Q Quantitative

arithmetic is meaningful

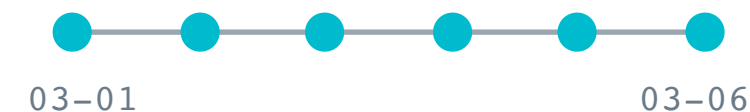


distance along the line
is meaningful

weight_kg, price

T Temporal

ordered and spaced



a position on a calendar —
order and spacing built in

delivered

The kind is not the storage type. A column stored as a number can be any of the four.

Check yourself

CHECK YOURSELF

- A postal code is stored as 79000 . Which kind is it?
- Can the same column be ordinal for one question and quantitative for another?
- Is a month name (Jan, Feb, Mar, ..., Dec) nominal, ordinal or temporal?
- The mean rating for Lviv is 4.5. What does that number assume?

3.2 · What one row means

Three questions about one week

THE OPERATIONS TEAM WANTS TO KNOW

- How much did we deliver on Saturday?
- What does an average delivery weigh?
- Which city receives the most parcels?

One row per day

DAILY — 7 ROWS

day	parcels	total_kg
Mon	180	402
Tue	165	371
Wed	172	388
Thu	190	447
Fri	240	566
Sat	262	611
Sun	155	344

THE OPERATIONS TEAM WANTS TO KNOW

- How much did we deliver on Saturday? ✓ 611 kg
- What does an average delivery weigh? ✓ $611 \div 262$
- Which city receives the most parcels? ✗ no city column

One row per delivery

DELIVERIES — SATURDAY ONLY, 262 ROWS

parcel	destination	weight_kg	courier
P-1041	Lviv	2.4	C-07
P-1042	Kyiv	1.8	C-07
P-1043	Odesa	11.0	C-12
P-1044	Kyiv	0.8	C-03
...	...	...	...

THE OPERATIONS TEAM WANTS TO KNOW

- How much did we deliver on Saturday? ✓ add up **weight_kg**
- What does an average delivery weigh? ✓ average **weight_kg**
- Which city receives the most parcels? ✓ count rows by **destination**

And questions the daily table cannot reach: the heaviest parcel, deliveries per courier, the share of the day going to one city.

Finer grain answers more questions



One row is one ...

Grain (granularity)

WHAT ONE ROW COUNTS

The level of detail at which a table records one observation. Before you compute anything from a table, you must be able to finish the sentence: one row is one ...

EXAMPLE

`daily` is one row per day. `deliveries` is one row per parcel. Read one grain as the other and every number you compute answers a different question from the one you asked.

Aggregate down, never up

A finer table can always be summarised into a coarser one. The detail a summary drops cannot be computed back.

What is the grain?

STORE_DAILY

store	date	receipts	revenue	avg_receipt
ST-1	2026-03-05	120	14400	120
ST-1	2026-03-06	95	11020	116
ST-2	2026-03-05	60	7800	130
ST-2	2026-03-06	80	9600	120

ANSWERABLE OR NOT?

- Total revenue for ST-1 over both days? ✓ add up revenue
- Average receipt for ST-1 over both days? ✓ from the totals, not the averages
- The single largest receipt? ✗ not at this grain
- What each customer bought? ✗ not at this grain

The average-of-averages trap

STORE_DAILY — ST-1 ONLY

store	date	receipts	revenue	avg_receipt
ST-1	2026-03-05	120	14400	120
ST-1	2026-03-06	95	11020	116

✗

MEAN THE AVG_RECEIPT COLUMN

$(120 + 116) \div 2$. The two days count equally, though one had 120 receipts and the other 95.

118.0

✓

REBUILD FROM THE TOTALS

$(14400 + 11020) \div (120 + 95)$. Every receipt counts once.

118.2

Here the gap is small because the two days have similar counts. The more unbalanced the groups, the bigger the gap, and you only learn its size by doing the check.

Check yourself

CHECK YOURSELF

- Finer grain answers more questions. Why would anyone store a summary at all?
- A daily summary table says Saturday was 611 kg. Adding up the individual deliveries gives 598 kg. Which number do you trust, and what do you do next?
- You are handed a table and no documentation. How do you work out its grain?

4 • Before your practice

What you need

Item	Why
A GitHub account	Practice notebooks are handed out and submitted through it
Git installed	Clone, commit, push
<code>uv</code> installed	Creates the Python environment from a locked file, identically on every machine

Full instructions are in the Week 1 section on Moodle, including a short setup check to run before your first practice. If it fails, bring the error message to the session.

Sources

McKinney, *Python for Data Analysis*, 3rd ed. — Ch. 1, Ch. 5.

McGregor, *Practical Python Data Wrangling and Data Quality* — Ch. 1, Ch. 3.

Wickham, "Tidy Data", *Journal of Statistical Software* 59(10).

Questions

What is the grain? — goals

GOALS

home	away	score	scorer	minute
Arsenal	Chelsea	3:1	Saka	12
Arsenal	Chelsea	3:1	Saka	47
Arsenal	Chelsea	3:1	Ødegaard	60
Arsenal	Chelsea	3:1	Palmer	80
Everton	Fulham	1:1	Keane	25
Everton	Fulham	1:1	Wilson	78

ANSWERABLE OR NOT?

- Who scored most?
- How many matches are in the table?
- Which team did each scorer play for?
- Which matches finished 0:0?

What is the grain? — sensors

READINGS

sensor_id	recorded_at	temp_c	humidity
S-04	2026-03-02 09:00:00	4.1	71
S-04	2026-03-02 09:15:00	4.3	70
S-04	2026-03-02 09:30:00	4.6	70
S-09	2026-03-02 09:00:00	6.8	63
S-09	2026-03-02 09:15:00	7.0	62

ANSWERABLE OR NOT?

- Warmest sensor on 2 March?
- Average temperature per hour?
- How long was S-04 above 5°C?
- How many sensors were offline?