Data Science Fundamentals

Concept revision through one student-support project

Mohammad Sabokrou

New Uzbekistan University

28 October 2026

The question and its measurement

Who may need support after the first four teaching weeks?

Population: the current entry cohort

Inputs: information available by the end of Week 4

Purpose: offer timely academic support

Login activity is a proxy for engagement

Could low platform use coexist with strong learning?

Population, frame and sample

A library survey can miss the students we most need to understand

Population: everyone our conclusion concerns

Sampling frame: the list or channel used to reach them

Sample: the people we actually observe

Selection bias can remain in a very large sample

Who is less likely to appear in the library survey?

Rows, joins and denominators

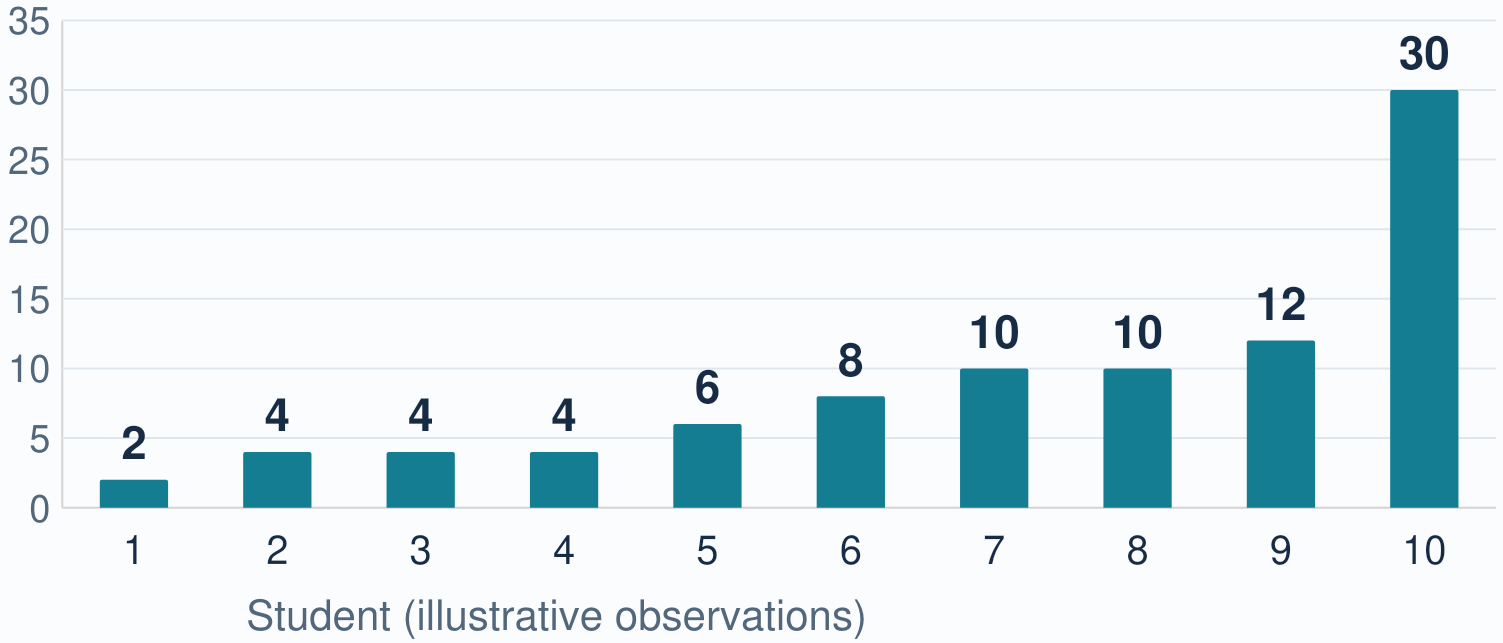
The same student can appear once, per course, or per week

Table	One row represents	Useful for
Students	One student	Counting people
Enrolments	One student-course registration	Course pass rates
Activity	One student in one week	Early activity history

A mean after a join can weight students by their number of records

Mean and median answer different questions

Illustrative activity counts: 2, 4, 4, 4, 6, 8, 10, 10, 12, 30



Mean = 9

Median = 7

Which summary best describes a typical student here?

Spread, quartiles and outlier flags

The same ten activity counts, summarized in different ways

Summary	Meaning in this example
Range: 30 – 2 = 28	Uses only the extremes
Q1 = 4, median = 7, Q3 = 10	The middle half lies between 4 and 10
IQR: 10 – 4 = 6	Spread of the middle half
Upper fence: 10 + 1.5 × 6 = 19	30 is flagged for investigation

An outlier flag alone does not justify deleting a record

Cleaning needs evidence

A valid-looking record can still describe the wrong person or event

Issue	Defensible action
Repeated student ID	Check the row unit and whether this is a real update
Attendance recorded as 140%	Verify definition, unit and original record
Different spellings of one program	Standardize against a documented category list
An unusually high activity count	Investigate before deleting or changing it

A cleaning log records the issue, action, reason and affected rows

Missingness changes what we can claim

A missing activity value does not establish zero activity

Zero: tracking worked and no activity occurred

Missing: the activity value is unknown

Not applicable: the measurement does not apply

Deleting or filling values can change group comparisons

Could students who need support be less likely to provide data?

Scaling changes the representation

Distance-based rules can be dominated by a variable’s numerical scale

Transformation	Purpose	What to remember
Standardization	Distance from the training mean, in SD units	Does not make data normal
Min–max scaling	Position within the training range	New values can fall outside 0–1
Log transformation	Compress a long right tail	Changes distances, not just units

Converting minutes to seconds should not silently decide who is “similar”

One fitted pipeline, reused for new students

Week 4’s example: early score 78, missing logins, program CS

Step	Rule learned from training data	New student
Impute	Training login median = 4	Missing logins become 4
Standardize	Score mean 70, SD 10; logins mean 4, SD 2	Score 0.8; logins 0
Encode	Category order: CS, Economics, Psychology	CS becomes [1, 0, 0]

The new student uses saved rules without redefining them

Event time and availability time

Prediction snapshot: the end of Week 4

Candidate input	Use at Week 4?	Reason
W1–W4 logins already visible	Yes	Recorded and accessible
W3 quiz score released in W5	No	The result arrives too late
Exam date announced in W1	Yes	The schedule is already known
Whole-semester attendance	No	Includes later observations

Information availability matters more than the event date

Training and held-out evaluation

Evaluation must reflect information available for new cases

Split before learning imputation and scaling rules

Fit the whole pipeline using training data only

Apply saved rules to held-out students

Low training error does not guarantee low future error

If the same student appears in both sets, what is the test really checking?

Patterns need an interpretation and a limit

Early activity and later outcomes may move together

A scatter plot reveals shape, clusters and unusual points

Pearson correlation summarizes a linear relationship

A strong curved relationship can have correlation near zero

Prior readiness may influence both activity and outcomes

Does an observed association establish that increasing logins improves learning?

Simpson’s paradox in the support project

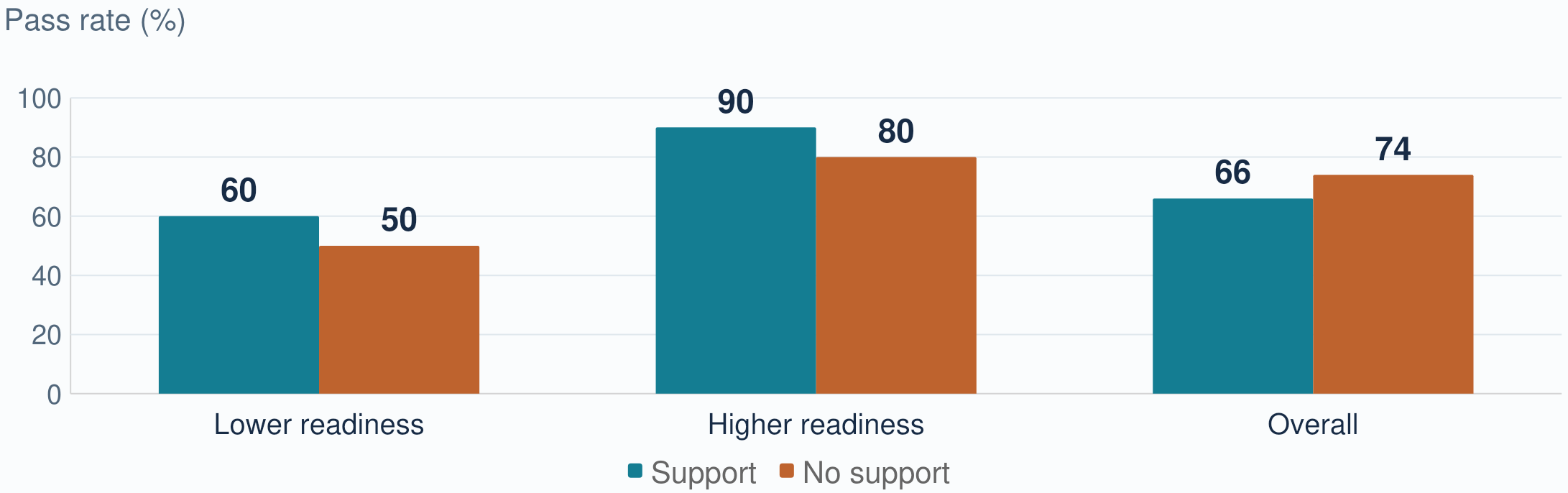
The illustrative 200-student example from Week 5

Baseline readiness	Support	No support
Lower readiness	48 / 80 = 60%	10 / 20 = 50%
Higher readiness	18 / 20 = 90%	64 / 80 = 80%
Overall	66 / 100 = 66%	74 / 100 = 74%

Support leads within both levels, but no support leads overall

Different group sizes create different weights

Rates need their denominators and group composition



Lower-readiness share: 80% with support, 20% without support

Chart choice follows the question

The support project requires several different views

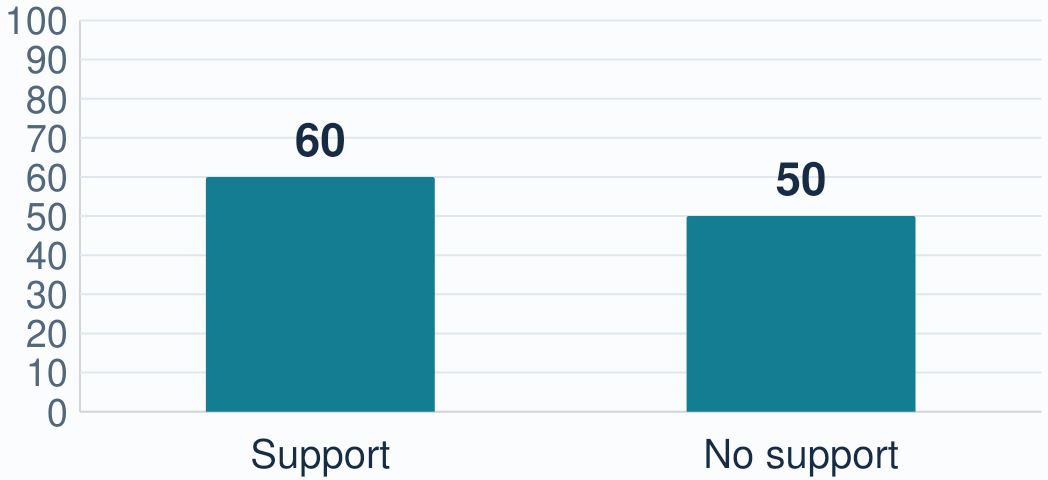
Question	Suitable view
How is student activity distributed?	Histogram or density curve
How do distributions differ across programs?	Box plots, with group sizes
How do early scores and final grades relate?	Scatter plot
How do pass proportions compare?	Grouped bars with denominators
How does activity change across weeks?	Line chart in time order

Histogram bins, box plots and correlations all hide some detail

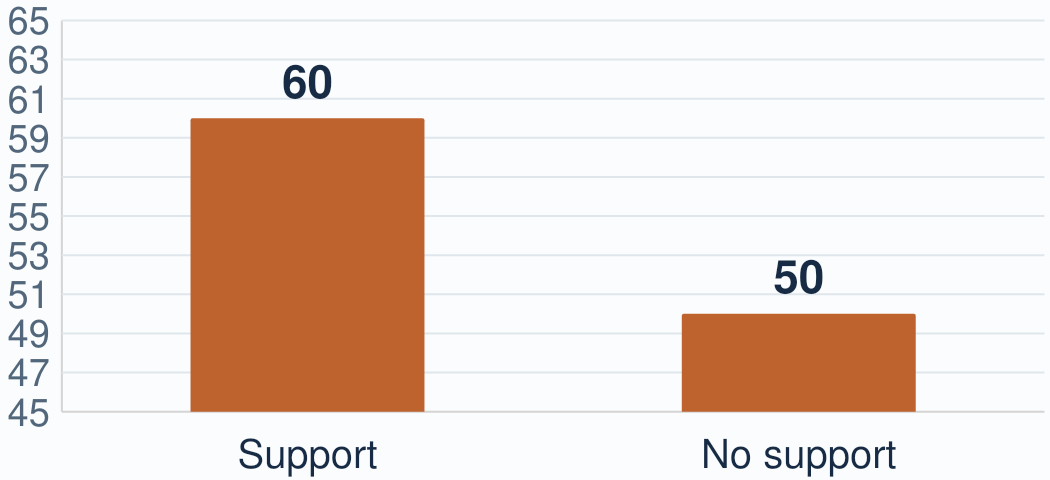
An honest chart shows the scale

The same lower-readiness pass rates: 60% with support and 50% without

Zero baseline



Truncated axis: 45–65%



Lower readiness: support 48/80; no support 10/20

The gap is 10 percentage points, or 20% relative to the 50% baseline

A conclusion needs context and care

A short reporting pattern for this support comparison

Observation: support has higher rates within each readiness level

Context: the readiness mix differs, reversing the overall comparison

Limit: this observational result does not establish a causal effect

Privacy: small filtered groups can reveal individual outcomes

A useful report names the population, period, denominator and limitation

Discussion: audit the whole project

An analyst proposes this workflow

Survey only library users and call the sample representative

Replace every missing activity value with zero

Fit a scaler on all students and use a late-released quiz score

Report only overall rates and claim that support causes harm

Which decisions would you change, and why?

Five questions to carry into every analysis

The important ideas from Weeks 1–6

What does one row measure, and who is missing?

Which summary preserves the information we need?

What changed during cleaning or preprocessing?

Was every input available at the prediction time?

Do the comparison and chart support the claimed conclusion?