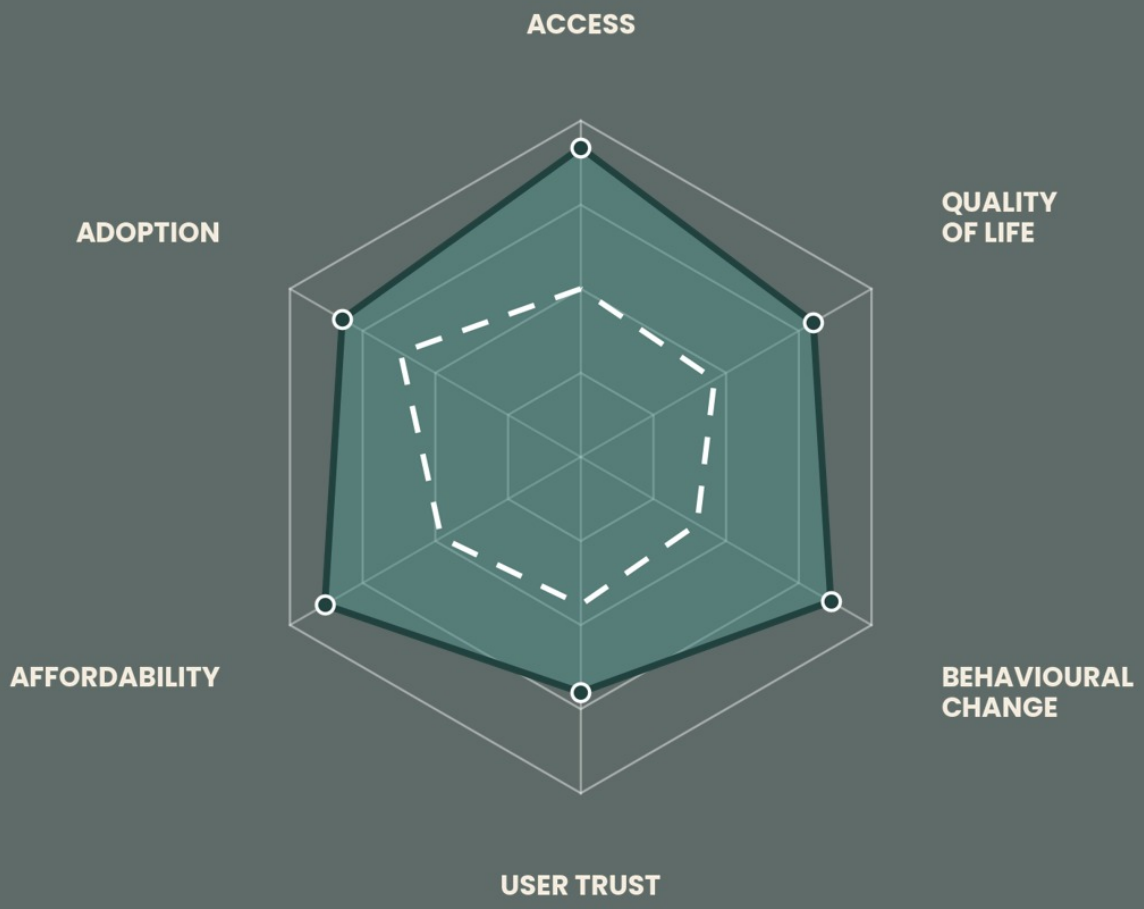
TRANSITIONS LAB



TRANSITIONS LAB · METHODOLOGY

Making Data Mean Something

How the Lab turns raw data into evidence a decision can rest on: cleaning, exploring, testing, and the discipline of not fooling yourself.

v1.0 · 2026 · transitionslab.org

START HERE

How to use this resource

This resource explains how the Lab analyses quantitative data: the pipeline from a messy dataset to a defensible finding, the tests that separate a real effect from noise, and the traps, correlation mistaken for cause, the misleading average, the tortured p-value, that turn analysis into fiction. It is a working guide, not a statistics textbook.

What this is

A working reference, not a marketing brochure. Read it in order, or jump to the section you need using the contents opposite.

Who it is for

Programme managers and funders who commission or read data analysis, researchers who want a disciplined workflow, and anyone who has been handed a chart and needs to know whether to trust it. It assumes numeracy, not statistical training.

How to get value from it

01 Learn the pipeline
Sections 1 to 4 walk the four stages every dataset goes through: clean, explore, test, explain. This is the backbone.

02 Understand the tests
Section 5 explains, in plain language, how the Lab decides whether an effect is real, and what a p-value does and does not mean.

03 Learn to spot the traps
Section 6 is the most useful part for a reader: the recurring ways data analysis misleads, and how to catch them.

04 See it applied
Section 7 works a real example end to end, and Section 8 shows how the numbers meet the fieldwork.

CONTENTS

What is inside

METHODOLOGY

What data analysis is for

The point of analysis is not a chart. It is a decision made more wisely than it would have been without the data. Everything in this guide serves that, and any analysis that does not change what someone does is decoration.

Data does not speak for itself. A spreadsheet of numbers is not evidence; it is raw material. Analysis is the disciplined process of turning that material into a claim you can defend, and, just as importantly, of knowing how far you can defend it. The best analysts are distinguished less by the sophistication of their methods than by their honesty about uncertainty.

The Lab runs every analysis through four stages, in order. Each depends on the one before it, and skipping or rushing an early stage quietly corrupts everything that follows.

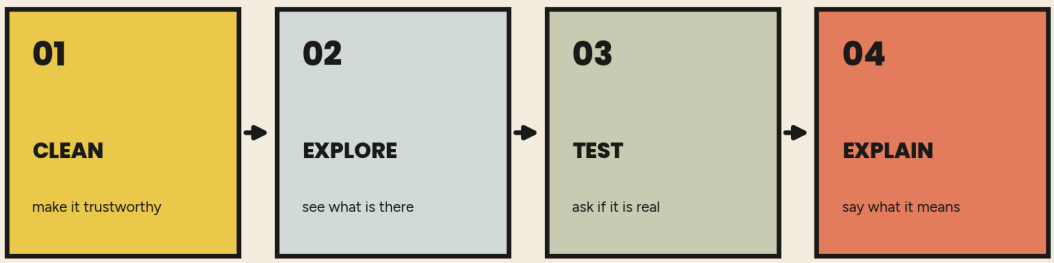


Figure 1. The analysis pipeline. Clean makes the data trustworthy; Explore reveals its shape; Test asks whether an effect is real; Explain says what it means for the decision. The order is not optional.

Most analytical disasters happen because someone jumped to the test, or to the chart, before the data was clean and understood. The stages that follow take each in turn.

FIRST, KNOW YOUR DATA

Not all numbers are the same

Before analysing anything, you have to know what kind of data you hold, because the type determines which analysis is legitimate. Treating one type as another is a quiet, common source of nonsense results.

Type	What it is	What you can and cannot do
Categorical	Named groups with no order: region, gender, crop type.	Count and compare proportions. You cannot average them, the 'mean region' is meaningless.
Ordinal	Ordered ranks without equal gaps: a 1-5 satisfaction scale.	Report medians and distributions. Averaging is common but treacherous: the gap from 1 to 2 may not equal the gap from 4 to 5.
Count	Whole-number tallies: trips per day, visits per week.	Often skewed; the average is easily distorted by a few high counts. Look at the distribution.
Continuous	Measured quantities on a real scale: income, distance, time.	The full toolkit applies, means, spreads, correlations, but only after checking the shape.

KEY POINT

The most common type error

Averaging an ordinal scale as if it were continuous. 'Average satisfaction is 3.7' treats the distance between 'unhappy' and 'neutral' as identical to the distance between 'happy' and 'very happy', which is rarely true. Report the distribution, and the Lab usually does both.

STAGE ONE

Clean: make the data trustworthy

The least glamorous stage decides the credibility of everything after it. Real-world data arrives messy, and analysis run on dirty data produces confident, precise, wrong answers.

Field data is never clean on arrival. Surveys have missing answers, duplicate entries, impossible values (an age of 200, a income of minus ten), and inconsistent categories ('F', 'female', 'Female' as three different groups). Before any analysis, the Lab audits and repairs the dataset, and, crucially, documents every decision made, so the cleaning itself can be reviewed.

Problem	What the Lab does
Missing values	Decide, and record, whether each is dropped, flagged, or imputed, and never silently. Why a value is missing often matters more than the gap itself.
Impossible values	Flag and investigate rather than delete. An age of 200 is an error; an income of zero may be the finding.
Duplicates	Identify and resolve, distinguishing genuine repeats from two real people with the same details.
Inconsistent categories	Standardise, so the same thing is counted as the same thing. This alone changes many headline numbers.
Outliers	Examine, never automatically remove. An outlier is sometimes an error and sometimes the most important case in the dataset.

KEY POINT

The rule that saves analyses

Every cleaning decision is documented and reversible. A finding that depends on quietly dropping inconvenient rows is not a finding; it is an artefact. If the cleaning cannot be shown, the result cannot be trusted.

STAGE TWO

Explore: see what is actually there

Before testing anything, look. Exploratory analysis is how you understand the shape of the data, and how you avoid the single most common error: trusting an average that hides everything that matters.

The average is the most over-used and under-examined number in analysis. It compresses a whole distribution into one figure, and when the distribution is skewed, or has two peaks, or a long tail, the average describes no one. A village where most earn little and a few earn a great deal has a ‘mean income’ that neither group would recognise.

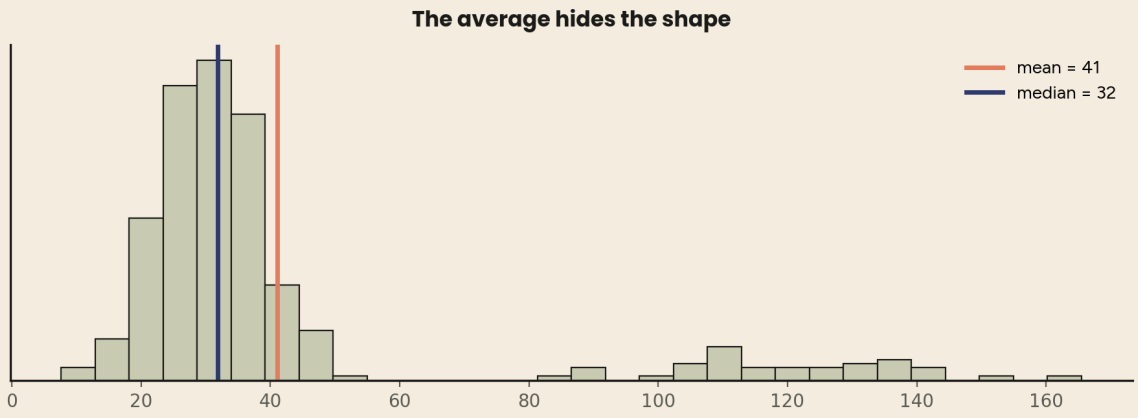


Figure 2. Why the shape matters. When a distribution has a long tail, the mean (pulled up by the few high values) and the median (the typical case) tell different stories. Reporting only the mean would misrepresent almost everyone.

Exploration means plotting the distribution, not just computing its average; comparing the mean and the median; looking for outliers, clusters, and gaps. It is where the analyst forms honest expectations before any formal test, and where surprises, often the real findings, first appear.

STAGE THREE

Test: is the effect real?

Once you can see an effect, the question is whether it is real or whether it could easily be noise. This is what statistical testing is for, and it is more modest than it is often made to sound.

Suppose adoption rose after an intervention. Before crediting the intervention, ask: could a rise this size have happened by chance, even if the intervention did nothing? A statistical test estimates how surprising the result would be in a world where there was no real effect. If it would be very surprising, we take the effect seriously. If it would be unremarkable, we do not.

Concept	In plain language
The null hypothesis	The sceptical starting point: assume there is no real effect, and see if the data can overturn it.
Statistical significance	A measure of how unlikely the observed result would be if there were truly no effect. Not a measure of how large or important the effect is.
Effect size	How big the difference actually is. A tiny, useless effect can be statistically significant in a large sample; this is why size matters as much as significance.
Confidence interval	The range the true value plausibly lies in. Almost always more informative than a single number or a yes/no verdict.

The Lab reports effect sizes and confidence intervals, not just significance verdicts. ‘Adoption rose 12 to 18 percentage points’ tells a decision-maker far more than ‘the result was significant ($p<0.05$)’, which says only that something happened, not how much.

A CLOSER LOOK

The p-value, honestly

No statistic is more used, or more misunderstood, than the p-value. Because it gates so many funding and publication decisions, it is worth stating plainly what it does and does not mean.

KEY POINT

What a p-value is

The probability of seeing a result at least this extreme if there were truly no effect. A small p-value means the data would be surprising in a no-effect world, so the no-effect assumption looks doubtful. That is all it means.

What a p-value is not:

- It is not the probability that the finding is true. A p-value of 0.05 does not mean there is a 95% chance the effect is real. It is a statement about the data under an assumption, not about the assumption itself.
- It is not a measure of importance. With a large enough sample, a trivial effect becomes 'significant'. Significance and importance are different questions.
- It is not a licence to stop thinking. $p<0.05$ is a convention, not a law of nature, and a result at $p=0.049$ is not meaningfully different from one at $p=0.051$.

KEY POINT

P-hacking: the trap to name

If you test enough relationships, some will cross the significance line by chance alone. Deciding what to test after seeing the data, or running many tests and reporting only the ones that 'worked', manufactures false findings. The Lab pre-specifies its main questions before analysis, and reports what it set out to test, not only what turned out well.

THE READER'S DEFENCE

How data analysis misleads

Most misleading analysis is not fraud; it is honest error, or motivated reasoning that stops short of a lie. These are the recurring patterns, and every reader of data should be able to spot them.

The most important one has its own name and its own page: mistaking correlation for causation.

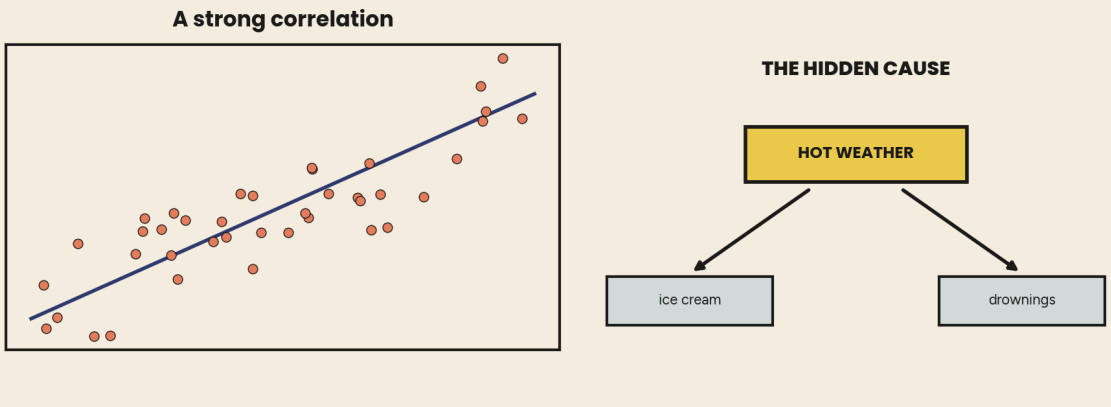


Figure 3. Correlation is not cause. Ice-cream sales and drownings rise together, not because one causes the other, but because a third factor, hot weather, drives both. The hidden common cause is called a confounder, and it is everywhere in real data.

Two things moving together is not evidence that one moves the other. They may share a hidden common cause (a confounder), the causation may run the opposite way, or the link may be coincidence. Establishing cause needs more than correlation: a comparison group, a natural experiment, or a carefully traced mechanism. See the Lab's Measurement Method guide on causal inference.

THE READER'S DEFENCE, CONTINUED

More traps to catch

Trap	What it looks like, and the defence
Survivorship bias	Analysing only the cases that remained (the users who stayed, the businesses that survived) and missing the ones that dropped out. Defence: find and study who is missing.
The misleading average	A single mean standing in for a skewed or two-peaked distribution. Defence: ask for the distribution, the median, and the range, not just the mean.
Cherry-picked baselines	Choosing a start point that flatters the trend. Defence: look at the whole series, not the convenient window.
Simpson's paradox	A trend that reverses when the data is split by group. Defence: check whether the headline holds within each subgroup, not just overall.
Overfitting	A model tuned so tightly to past data that it fails on new data. Defence: test on data the model has never seen.
False precision	Reporting '47.3%' from a sample of thirty, implying an accuracy the data cannot support. Defence: match the precision to the uncertainty.

The common thread: each trap makes a result look stronger or cleaner than the evidence warrants. The defence is always the same, ask what the analysis leaves out, and whether the finding survives once you put it back.

A WORKED ANALYSIS

One dataset, from raw file to finding

To make the pipeline concrete, here is a real-shaped analysis end to end: a programme claims its electric-mobility financing raised rider incomes. Does the data support it?

WORKED EXAMPLE

From messy file to defensible claim

Clean. 1,204 rider records. 38 have impossible incomes (negative or absurd), flagged and excluded with a note. 112 have missing endline income, kept but marked; we check later whether they differ systematically from the rest (they do: they are more likely to have dropped out, which matters).

Explore. Income is right-skewed: most riders gain modestly, a few gain a lot. The mean rise (KSh 4,100) overstates the typical rider; the median rise (KSh 2,600) is the honest headline. We report both.

Test. Comparing endline to baseline, the median rise is unlikely to be chance (the confidence interval, KSh 1,900 to 3,300, sits well clear of zero). But baseline-vs-endline alone cannot rule out that incomes rose for everyone that year.

The honest limit. Without a comparison group of non-adopters, we cannot fully separate the programme's effect from a general trend. We say so plainly, and quantify what we can defend rather than what we would like to claim.

The finding. ‘Adopting riders saw a median income rise of KSh 2,600 (CI 1,900 to 3,300) over the period. The pattern is consistent with a real programme effect, but a comparison group would be needed to confirm causation.’

THE COMPLETE PICTURE

Where numbers meet fieldwork

Quantitative analysis tells you what happened and how much. It rarely tells you why, or what it meant. That is why the Lab never reads numbers alone.

In the worked example, the data showed a median income rise but could not explain the riders who gained nothing, or the ones who dropped out. Only the interviews did that: the dropouts had lost hours to unreliable battery-swapping, and the gains concentrated among riders on busy routes. The number is the skeleton; the fieldwork is the flesh.

What the numbers gave

The size and spread of the income effect, who gained most, and the honest limit of what baseline-endline can prove. Comparable, scalable, defensible.

What the fieldwork gave

Why the dropouts left, why the gains concentrated, and what riders would fix first. The mechanism behind the number, and the friction to act on.

WORKING WITH TRANSITIONS LAB

Analysis you can defend

The Lab handles the whole chain: clean data, honest analysis, and a written reading of what it means, paired with the fieldwork that explains it.

IN THIS FREE RESOURCE	IN AN ENGAGEMENT, WE GO FURTHER
Clean and document the data	› A dataset you and an auditor can both trust, with every decision recorded.
Test honestly, report the uncertainty	› Effect sizes and confidence intervals, not significance theatre.
Explain with the fieldwork	› The number and the reason behind it, read together.

The result is analysis a funder can rely on and a team can act on, honest about what it shows and what it cannot.

If a decision in your work turns on evidence of what a technology or programme actually does to real people, we are glad to talk it through, with no obligation. transitionslab.org

PRESENTING RESULTS

Charts that tell the truth

The last stage of analysis is communication, and it has its own ethics. The same honest finding can be shown fairly or made to mislead, and the choices are often invisible to the reader. The Lab holds to a few rules.

Rule	Why
Start bar-chart axes at zero	A truncated axis exaggerates small differences into dramatic ones. The most common way an honest number is made to lie.
Show the uncertainty	Error bars or a stated range tell the reader how firm the finding is. A bare bar implies a precision the data may not have.
Match precision to the sample	'47%' from thirty people should be 'about half'. False decimals imply false accuracy.
Show the distribution, not just the average	Where the shape matters, a single bar hides it. A simple spread or histogram is more honest.
Label plainly, and note what is excluded	A chart that quietly drops the dropouts, or the non-users, is telling a partial truth. Say what is not shown.

A chart is an argument. Every design choice, the axis, the grouping, the colour, nudges the reader toward a conclusion. The Lab's standard is that the nudge must match the evidence: the clearest honest reading, never the most flattering one.

REFERENCES

Sources and further reading

This guide draws on the applied-statistics and data-analysis literatures, chosen for clarity and honesty about uncertainty rather than technical depth.

Tukey, J. W. (1977). Exploratory Data Analysis. Reading, MA: Addison-Wesley.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. The American Statistician, 70(2), 129-133.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. American Scientist, 102(6), 460-465.

Reinhart, A. (2015). Statistics Done Wrong: The Woefully Complete Guide. San Francisco: No Starch Press.

Huff, D. (1954). How to Lie with Statistics. New York: W. W. Norton.

Wickham, H. (2014). Tidy data. Journal of Statistical Software, 59(10), 1-23.

Ioannidis, J. P. A. (2005). Why most published research findings are false. PLoS Medicine, 2(8), e124.

Angrist, J. D., & Pischke, J.-S. (2009). Mostly Harmless Econometrics. Princeton University Press.

KEY POINT

Read alongside

This guide covers the analysis of data once gathered. For how the Lab decides what to measure and how to gather it, see the Measurement Method guide; for the depth work that explains the numbers, see the In-Depth Interview Guide.