Facilitator guide: Module 13, Data ethics, quality, and coverage bias

A teacher-facing companion to Module 13. The module is the full lesson; this guide adds the facilitation layer: a preparation checklist, timing cues, guidance for the discussion questions, and notes on what to emphasize and where students get stuck. Slides and a printable version are below.

At a glance

- **Length:** 90 minutes. A 60 minute and a 120 minute variant are noted under Teaching notes.
- **Level:** undergraduates and graduate students across data science, business, and social science programs. No finance or programming background is assumed.
- **Access:** entirely no-code, on the sandbox at sandbox.altfndata.com. No API key is needed. Students self-register with a work or school email and are approved automatically.
- **Goal of the session:** students leave able to recognize the recency under-ingestion pattern as a coverage-bias artifact rather than a market signal, and to draft a specific disclosure statement for a claim built on this data.

Before class

- Register your own sandbox account and run Query 1 through 3 for a chosen category, so you have real quarter and vendor counts to point to.
- Prepare or print the short handout defining coverage bias, point-in-time correctness, and disclosure, for reference during the small-group exercise.
- Open the coverage browser once beforehand and note which vendors are heavily versus lightly represented for your chosen category, so you can point to real numbers live.
- Capture a screenshot of the Query 1 result in case the sandbox is slow or unavailable during class.
- Decide which category you will demo with and have a backup ready in case the first shows too few rows to make the recency drop-off legible.

Timed agenda with cues

Time	Segment	What to do	Watch for
0 to 15	Introduce the central question	Pose how you would know if a pattern in the data is real or an artifact of how it was collected, and preview the recency case study.	Students assuming any dip must be a genuine market signal. Hold that instinct for the demo rather than resolving it now.
15 to 25	Sandbox orientation	Confirm students can open a data table and locate sale_date, status, vendor, and usd_price_decimal in the data dictionary.	Registration email typos. Have a backup shared account ready in case one student cannot get in.
25 to 45	Guided demo, counts by quarter	Run Query 1 and have students read the last two or three bars aloud. Explain the recency under-ingestion pattern directly.	Students reading the most recent quarters as evidence of a real decline. Redirect to the stable, longer-run shape.

Time	Segment	What to do	Watch for
45 to 60	Guided demo, counts by vendor	Run Query 2 and discuss uneven vendor coverage as a second, independent source of coverage bias.	Students conflating "the market" with the specific houses actually captured in the table.
60 to 75	Small-group exercise	Each group picks a different category, reproduces both counts, and drafts a one-paragraph disclosure statement for a claim built on that category.	Disclosures vague enough to apply to any dataset. Push groups toward specifics: which quarters, which vendors.
75 to 85	Class discussion	Groups share their disclosure statements. The class critiques whether each is specific enough to be useful to a reader.	Groups accepting a vague disclosure without pushback. Model one pointed follow-up question yourself first.
85 to 90	Wrap-up and homework	Restate the one-sentence takeaway and hand out the homework.	Leave 2 minutes for the homework logistics, not zero.

Guidance for the discussion questions

Use these as the "what to listen for" behind each question in the module. They are talking points, not a graded key. The recency under-ingestion caveat is the central case study running through this guidance.

1. **Distinguishing an ingestion artifact from a genuine decline.** Listen for: compare the drop-off shape across several categories and vendors at once. If the same recent-quarter dip appears broadly and fades once you look further back than 2 to 3 quarters, that is the signature of ongoing ingestion rather than an independent, category-specific demand shock. Point students back to Query 1 as exactly this test.
2. **Why a vague footnote is not a real disclosure.** Listen for: "may be incomplete" applies to any dataset ever built and gives a reader no way to judge how much to discount a figure. A useful disclosure names which quarters are affected, roughly how understated the count likely is, and points to the stable longer-run figure as the trustworthy comparison instead.
3. **An underrepresented house and which claims are most distorted.** Listen for: claims about "the market" as a whole, or about total volume or share, are most distorted by uneven vendor coverage. Claims scoped explicitly to the vendors actually captured, such as a within-house figure, are relatively unaffected since they never claimed broader coverage in the first place.
4. **Why dropping unsold lots changes sell-through, and what else breaks.** Listen for: sell-through is sold over offered, so removing unsold rows collapses the denominator and pushes the figure toward 100 percent by construction. Any calculation that needs an honest count of everything offered, not only what sold, breaks the same way, including demand comparisons across categories.
5. **Point-in-time correctness and sale_date.** Listen for: sale_date reflects when the transaction actually happened, while a retrieval or ingestion date reflects when the record was added to the dataset. A claim about "what happened last quarter" built on retrieval date instead would misattribute older transactions to the wrong period and compound the recency confusion rather than resolve it.
6. **Row-level accuracy versus a misleading aggregate claim.** Listen for: every individual row can be a true, verified record, and an aggregate built from an incomplete or unevenly sampled set of those rows can still misrepresent the whole. The analyst's responsibility is to check how the aggregate was constructed, not just whether each row is true, before making a claim.
7. **Explaining the recency caveat to a non-technical audience.** Listen for: a good explanation names the specific pattern, newest data still catching up, without dismissing it as trivial or implying the whole dataset is unreliable. Concrete language such as "the last quarter or two undercounts because new records are still

being added" works better than a soft hedge like "may."

8. **Limitation versus untrustworthy.** Listen for: every dataset has limitations, and naming them precisely is what makes a dataset usable. A dataset becomes untrustworthy only if its limitations are hidden or misrepresented. The distinction matters because it is exactly what a disclosure statement exists to convey, and confusing the two produces either false confidence or unwarranted rejection of otherwise good data.

Teaching notes

- **Most common misconception:** students read the recency dip as declining market demand. Run Query 1 live across two or three different categories to show the same pattern recurring broadly, which is structural, not a category-specific demand story.
- **Second misconception:** treating unsold lots as noise to clean out of a query. Recompute sell-through with and without unsold rows side by side so students see the denominator collapse directly.
- **60 minute variant:** drop the small-group disclosure-drafting exercise; keep the two guided demos and hold a shorter, instructor-led discussion of what Query 1 and Query 2 show.
- **120 minute variant:** after groups share their disclosure statements, have each group trade with another group for a round of peer critique, then revise before a final class discussion.
- **If the sandbox is slow or blocked on the room network:** fall back to the pre-run screenshots of Query 1 through 3 captured in preparation, and have students draft their disclosure statements from those instead of live results.

For the full lesson content, queries, and homework, see Module 13. Questions or a class API key: info@altfndata.com.