

Data ethics, quality, and coverage bias

This session teaches students to read a real dataset critically rather than taking any figure it produces at face value. The central case study is a coverage bias built into how this dataset, like most transaction datasets assembled from many independent sources, actually gets built:

What today looks like

- 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.

OBJECTIVES

By the end of class students can

- Define coverage bias and explain, using the recency under-ingestion pattern in this dataset, how an ingestion artifact can be mistaken for a real trend if a dataset's construction process is not understood.
- Query row counts by quarter and by vendor in the sandbox to confirm the recency pattern directly, rather than accepting the caveat on faith.
- Explain why unsold lots must be retained and correctly labeled rather than dropped, and describe what would be distorted, such as sell-through, if they were excluded.
- Describe point-in-time correctness and explain why a query that ignores `sale_date` can produce a claim that misrepresents when a transaction actually happened.
- Distinguish a claim a dataset can support from one it cannot, and draft a disclosure statement appropriate to a specific claim.
- Identify at least three sources of coverage bias in this dataset beyond recency, such as uneven vendor participation or category-specific gaps, and explain how each could mislead an unqualified reader.

How the session runs

- 0 to 15 min: Introduce the session's central question, how would you know if a pattern in the data is real or an artifact of how the data was collected, and preview the recency case study.
- 15 to 25 min: Sandbox orientation. Confirm students can open a data table and locate `sale_date`, `status`, `vendor`, and `usd_price_decimal` in the data dictionary.
- 25 to 45 min: Guided demo, count rows by quarter for a single category and observe the recency drop-off directly, then discuss why it reflects ingestion timing rather than a market decline.
- 45 to 60 min: Guided demo, count rows by vendor for the same category and discuss uneven vendor coverage as a second, independent source of coverage bias.
- 60 to 75 min: Small-group exercise, each group picks a different category, reproduces both counts, and drafts a one-paragraph disclosure statement they would attach to a claim built on that category's data.
- 75 to 85 min: Class discussion, groups share their disclosure statements and the class critiques whether each one is specific enough to be useful to a reader.
- 85 to 90 min: Wrap-up and homework assignment.

The in-class walkthrough

- Open sandbox.altfndata.com, sign in, and select a data table, for example the fine art data, from the SQL editor dropdown.
- Open the coverage browser tab and look at how several vendors are represented in the vendor field, noting that participation is not evenly distributed across houses.
- Return to the SQL editor and run the first guided query, sold-lot counts by quarter, and have students read the last two or three bars or rows aloud. Ask what they notice about the most recent quarters compared to the ones before them.
- Explain the recency under-ingestion pattern directly: the newest quarters are still being ingested at any given point in time, so the drop-off in the most recent quarters is a data collection artifact, not evidence that fewer transactions occurred or that a market cooled. Emphasize that this is exactly the kind of pattern a careless reader could mistake for a real trend, and exactly why any time-bucketed claim in this course has been built on stable, longer-run figures rather than the newest quarters.
- Run the second guided query, sold-lot counts by vendor, and discuss how unevenly represented vendors are, and what that means for any claim based on "the market" as a whole rather than on the specific houses actually captured.
- Introduce point-in-time correctness using the status field: explain that an unsold lot is retained in the data and marked unsold, and ask what would happen to a sell-through calculation, or to any claim about how much of a category found a buyer, if unsold lots were silently dropped instead.

The core query

- Run this in the sandbox SQL editor:

```
- SELECT date_trunc('quarter', CAST(sale_date AS date)) AS quarter,  
-          COUNT(*) AS sold_lots  
- FROM all_fine_art_data  
- WHERE status = 'sold'  
- GROUP BY date_trunc('quarter', CAST(sale_date AS date))  
- ORDER BY quarter DESC  
- LIMIT 12;
```

DISCUSS

Questions for the room

- If the most recent two quarters in a chart show a steep drop-off, what evidence would distinguish an ingestion artifact from a genuine decline in market activity, and how would you gather that evidence using only this dataset?
- Why is it not enough to simply state "the newest data may be incomplete" in a footnote? What would a more specific and useful disclosure look like?
- If a house's lots are underrepresented in this dataset relative to its actual share of the market, what kinds of claims about "the market" would be most distorted by that gap, and which claims would be relatively unaffected?
- Why does treating an unsold lot as if it never existed change a sell-through calculation, and what other calculations in this dataset would be similarly distorted by dropping unsold rows?
- What does point-in-time correctness mean in the context of `sale_date`, and what claim would be wrong if it used a transaction's retrieval date instead of its actual sale date?

This week's assignment

- Each student selects one category table and one claim they might want to make about it, for example "which auction house sells the most in this category" or "how has demand for this category changed over time." The student runs the three guided queries (or their equivalents) for their chosen category, writes a one-page memo stating the claim, the evidence for it, and a disclosure section that specifically addresses recency under-ingestion, vendor coverage, and correct treatment of unsold lots as they apply to that claim.
- The memo must include the SQL queries used as an appendix and the exported result sets.
- Grading criteria:
- correct construction and reading of the three queries (25 percent), accuracy and specificity of the coverage bias analysis (35 percent), quality and honesty of the disclosure section (30 percent), and clarity of the memo (10 percent).

NEXT

Wrapping up

- Restate the one-sentence takeaway before students leave.
- Point to the facilitator guide for discussion guidance and teaching notes.
- Questions or a class API key: info@altfndata.com