

Study sheet: Module 13, 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: the newest quarters are still being ingested at the time any query is run, so a naive comparison of recent-quarter counts against older ones will show a decline that looks like falling activity but is actually an artifact of incomplete ingestion, not a market signal. Students confirm this pattern themselves by counting rows per quarter and per vendor in the sandbox, then use what they find to develop a broader set of habits for responsible data use: disclosing what a dataset does not see, treating unsold lots correctly rather than discarding them, respecting point-in-time correctness, and stating a claim's limitations alongside the claim itself. The session's goal is not to teach students to distrust data, but to teach them to read it the way a careful analyst does.

What you should be able to do

1. 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.
2. Query row counts by quarter and by vendor in the sandbox to confirm the recency pattern directly, rather than accepting the caveat on faith.
3. 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.
4. 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.
5. Distinguish a claim a dataset can support from one it cannot, and draft a disclosure statement appropriate to a specific claim.
6. 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.

Key terms

- **Coverage bias:** a systematic gap between what a dataset captures and what it claims or is assumed to represent, arising from how the data was collected rather than from the underlying reality.
- **Recency under-ingestion:** the specific coverage bias in this dataset in which the newest quarters have fewer records than they eventually will, because ingestion of new transactions is ongoing.
- **Point-in-time correctness:** the property that a record's date fields accurately reflect when the underlying event, here a sale, actually occurred, as distinct from when it was added to the dataset.
- **Unsold lot:** a lot that was offered at auction but did not meet its reserve price and therefore did not sell, retained in the data and marked with a status of unsold rather than removed.
- **Sell-through:** the share of offered lots that find a buyer, a figure that requires both sold and unsold lots to be present and correctly labeled to compute honestly.
- **Disclosure:** an explicit statement, attached to a claim, of the limitations in the underlying data that a reader should weigh before accepting the claim.

- **Survivorship-style gap:** a general category of coverage bias in which the data available for analysis systematically omits certain cases, distinct from but related to the recency gap studied in this session.
- **Vendor coverage:** the degree to which different auction houses or marketplaces are represented in a dataset, which need not match their actual share of real-world transaction volume.

Core sandbox query

The query the module is built around. Run it, then change the brand or category and compare.

```
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;
```

Full lesson, more queries, and the homework are in Module 13 at academy.altfndata.com. Questions: info@altfndata.com.