

Problem set 13: Data ethics, quality, and coverage bias

Module 13. Total: 100 points. Format: no-code sandbox.

Before you start

This problem set asks you to confirm, with your own queries, a coverage bias built into how this dataset is actually assembled: the newest quarters are still being ingested at the time any query runs, so a naive comparison of recent-quarter counts against older ones can look like a market decline when it is really an artifact of incomplete ingestion. Choose one category table (for example `all_watches_data`, `all_jewels_gems_data`, `all_handbags_data`, or `all_fine_art_data`) and one claim you might want to make about it, such as which auction house sells the most in that category, or how demand for that category has changed over time. Use that same category and claim throughout this problem set.

Problem 1 (15 points): Coverage bias and recency under-ingestion

In four to five sentences, define coverage bias in general terms, then explain the specific recency under-ingestion pattern in this dataset, and describe how an ingestion artifact like this one could be mistaken for a real market trend by a reader who does not understand how the dataset is built.

Problem 2 (20 points): Confirm the recency drop-off yourself

Write a SQL query against your chosen category table that counts sold lots by quarter, ordered most recent first.

```
-- your query here
```

In two to three sentences, describe what you observe in the most recent two or three quarters compared to the ones before them, and state explicitly why that drop-off should not be read as evidence of falling market activity.

Problem 3 (20 points): Vendor coverage as a second, independent source of bias

Write a SQL query against the same table that counts sold lots by vendor, ordered highest first.

```
-- your query here
```

In three to four sentences, explain what this result tells you about how evenly, or unevenly, different auction houses are represented in the data, and describe one specific claim about "the market" as a whole that this unevenness would distort, versus one claim that would be relatively unaffected by it.

Problem 4 (20 points): Point-in-time correctness and unsold lots

Write a SQL query against the same table that compares sold-lot counts to total offered-lot counts by quarter, computing sell-through for each quarter.

-- your query here

In three to four sentences, explain why an unsold lot must be retained in the data and correctly labeled rather than dropped, what would happen to your sell-through figures if unsold lots were silently excluded, and whether the same recency caution from Problem 2 applies to a quarter's sell-through figure and why.

Problem 5 (25 points): A disclosure statement and a claims audit

Using your results from Problems 2 through 4, write a disclosure statement, 200 to 300 words, that you would attach to your chosen claim about your chosen category. The statement must specifically address recency under-ingestion, vendor coverage, and the correct treatment of unsold lots as they apply to your specific claim, not as generic caveats that could apply to any dataset. Then, in a short list, state one version of your claim that this data can support, and one version of a related claim, using the same data, that it cannot support, and explain the difference in one sentence each.

Submission. Turn in this file with your SQL, disclosure statement, and claims audit filled in.