

Data science for art market research

This session teaches the quantitative methods behind rigorous art market research, applied to real auction transaction data rather than to a textbook example. Where Module 7 treated the art market as a business and built the artifacts a saleroom or a gallery director reads, this module steps back to the method itself:

What today looks like

- Length: 90 minutes.
- Level: graduate art business or analysts, comfort with numbers and tables assumed, no heavy math and no statistics prerequisite beyond the idea of a median.
- Access: the no-code sandbox at sandbox.altfndata.com is the primary path for the entire session. The optional Python and API extension needs a class API key from your instructor account, requested from info@altfndata.com.
- Goal of the session: students leave able to build a defensible comparable set, construct and honestly read an artist-level market index, pull demand signals that do not depend on price, and run a cohort comparison across artists without mistaking noise for a finding.

OBJECTIVES

By the end of class students can

- Define a comparable set for an art market question, explain why it is the atomic unit of the analysis, and state the filters (artist, date range, status) that constitute one.
- Construct an artist-level market index from the median of realized price by period, and explain why the index tracks method and the changing composition of what sold in each period, not pure appreciation of the artist's work.
- Apply hedonic thinking, price as a function of observable characteristics, honestly against a dataset with few documented characteristic fields, using a keyword drawn from item_title as a rough and imperfect proxy rather than a true attribute.
- Read demand signals that do not depend on price at all, specifically sell-through and the bought-in (unsold) rate, and explain what each adds beyond a price index.
- Run a cohort or comparative analysis across several artists at once with a minimum-lot threshold, and explain why the threshold exists and what it protects against.
- Practice point-in-time discipline, using sale_date rather than any ingestion or load date to define a period, and state the recency and secondary-market-only caveats that qualify every result in this session.

How the session runs

- 0 to 12 min: What makes art market research rigorous. Introduce the comparable set as the atomic unit of the analysis and preview the four methods the class will build: an index, a hedonic-style split, demand signals, and a cohort comparison.
- 12 to 22 min: Sandbox and schema orientation. Confirm students can open the fine art data table and locate designer, item_title, sale_date, usd_price_decimal, sale_estimates_high_usd_price, status, and vendor in the schema browser, and confirm one artist's lot count is large enough to support an index.
- 22 to 42 min: Guided demo, build an artist-level market index from median realized price by year, with lot count shown alongside every period.
- 42 to 55 min: Guided demo, demand signals independent of price: sell-through and bought-in rate for the same artist, plus a rough hedonic split using an item_title keyword.
- 55 to 70 min: Guided demo, cohort comparison across several artists at once with a minimum-lot threshold, read as a comparative table rather than one artist studied alone.
- 70 to 82 min: Small-group exercise, students pick a different artist or a small cohort and reproduce the index and the demand signals, then flag which of their own recent periods should be treated with caution.
- 82 to 90 min: Class discussion and wrap-up, groups report what their index and demand signals show, and the instructor assigns the homework.

The in-class walkthrough

- Open sandbox.altfndata.com, sign in, and select the fine art data table from the SQL editor dropdown.
- Open the coverage browser tab and confirm how a well-known artist is written in the designer field, and confirm roughly how many sold lots that artist has. An artist with too few lots cannot support a reliable index or a comparative table.
- Return to the SQL editor and run the first guided query, the artist-level market index, which groups median realized price by year alongside the count of lots sold that year.
- Read the index aloud with the class and point to the count column first. Explain that the index moves for two reasons at once, the mix of works that happened to sell that year and any genuine shift in the artist's market, and that the count column is how a reader tells the two apart.
- State the recency caveat before anyone reads the newest one or two periods as a signal: those periods are still being ingested, lot counts there are understated, and a median computed on an incomplete period is not comparable to a median computed on a complete one. Treat the newest one or two periods as provisional, not as the current trend.
- Run the second guided query, sell-through and bought-in rate for the same artist, and explain that these numbers depend on nothing about price at all, only on how many offered lots found a buyer, so they are a useful check against a price index that might be moved by a handful of results.

The core query

- Run this in the sandbox SQL editor:

```
- SELECT DATE_TRUNC('year', sale_date) AS period,  
-         approx_percentile(usd_price_decimal, 0.5) AS median_realized_usd,  
-         COUNT(*) AS sold_lots  
- FROM all_fine_art_data  
- WHERE status = 'sold'  
-       AND designer LIKE '%Warhol%'  
- GROUP BY DATE_TRUNC('year', sale_date)  
- ORDER BY period;
```

Questions for the room

- A comparable set is defined by an artist filter, a date range, and a status filter. What happens to an index or a demand signal if any one of those three is loosely defined, and how would you know it had gone wrong?
- The market index moves from one year to the next. What are the two distinct reasons it can move, and why does showing the lot count next to the median help a reader tell them apart?
- Why does an index built from median realized price measure method and the composition of what sold, and not the pure appreciation of an artist's work, and what would it take to separate the two?
- Sell-through and bought-in rate depend on nothing about price. What can these two signals tell you about an artist's market that a price index cannot, and what can they not tell you?
- The hedonic-style split in Query 3 uses a keyword from item_title as a proxy for medium. What kinds of lots would this keyword approach misclassify, and how would you check how much error it introduces?

This week's assignment

- Each student selects one artist represented in the fine art data with at least 150 sold lots and produces a short methods memo, as if for a research director reviewing a junior analyst's first market study.
- The memo defines the comparable set used, builds the artist's market index by year with lot counts shown, reports sell-through and bought-in rate, attempts the hedonic-style keyword split and states its limitations, and places the artist in a cohort table alongside at least two other artists using the minimum-lot threshold from class.
- The memo must include the SQL queries used as an appendix, explicitly flag which of the artist's own recent periods are provisional due to ongoing ingestion, and close with a limitations section addressing the secondary-market-only nature of the data and the fact that the median index reflects composition and method as much as it reflects the artist's market.
- Grading criteria:
- correct construction of the comparable set and the index with lot counts shown (25 percent), correct computation and honest interpretation of the demand signals (25 percent), reasonable and appropriately caveated use of the hedonic-style keyword split (20 percent), quality and rigor of the cohort comparison (15 percent), and honest treatment of the recency and secondary-market-only limitations (15 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