ALT/FNDATA ACADEMY · STUDENT STUDY SHEET

Study sheet: Module 16, 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: how a researcher builds a defensible comparable set, constructs an artist-level market index and reads it honestly, looks for demand signals that do not depend on price at all, and compares several artists at once without mistaking noise for a finding. The dataset records what buyers actually paid at more than 100 auction houses with history back to the late 1990s, and every method in this session is built entirely from documented fields, so students leave able to defend each step of an analysis, including its limits, to a skeptical reader. This is the art-market-specific application of the program's data-science modules on SQL, visualization, machine learning, and time series, not a repeat of them.

What you should be able to do

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. (Extension) Retrieve one artist's sold-lot records programmatically via the production API and build the median-by-year index client-side in pandas, reproducing in code what the SQL editor computed server-side.

Key terms

- **Comparable set:** the group of lots, defined by an artist filter, a date range, and a status filter, that forms the atomic unit of an art market analysis.
- **Market index:** a period-by-period summary statistic, here the median of realized price, used to track a market's movement over time.
- **Composition effect:** the change in an index caused by a shift in which works happened to sell in a given period, rather than by a genuine change in the market.

- **Hedonic thinking:** treating price as a function of observable characteristics, applied here loosely because few characteristic fields are documented beyond item_title text.
- **Demand signal:** a measure of market activity, such as sell-through or bought-in rate, that does not depend on the price level at all.
- **Bought-in rate:** the share of offered lots that failed to find a buyer and were returned to the consignor, the complement of sell-through.
- **Cohort analysis:** a comparison of a market measure across several artists or groups at once, read side by side rather than one at a time.
- **Minimum-lot threshold:** a HAVING clause requirement that a group have at least a set number of lots before it appears in a comparative table, protecting against conclusions drawn from too few observations.
- **Point-in-time correctness:** defining a period strictly by sale_date, never by any date the record happened to be loaded or ingested.

Core sandbox query

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

```
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;
```

Common mistakes to avoid

1. Reading a rising median index as proof that an artist's market is appreciating, without checking whether the lot count for that period shows the mix of works has simply changed.
2. Treating the newest one or two periods in an index as the current trend, when those periods are still being ingested and their lot counts and medians are not yet complete.
3. Filtering by any date field other than sale_date to define a period, which breaks point-in-time correctness and can silently mix periods together.
4. Treating an item_title keyword match as a true characteristic field, rather than an imperfect proxy that will miss or misclassify some lots.
5. Building a cohort table without a minimum-lot threshold, letting a low-volume artist's noisy figure sit next to a well-supported one as if the two were equally reliable.
6. Reading sell-through or bought-in rate on its own as a complete demand picture, without also considering that a low sell-through can reflect poor consignment quality rather than weak demand.
7. Drawing a conclusion about an artist's overall market from auction data alone, without noting that the analysis is silent on primary or gallery sales.

Full lesson, more queries, and the homework are in Module 16 at academy.altfndata.com. Questions: info@altfndata.com.