Audience Validation Framework

Overview

This document describes the audience validation approach used to measure and compare the quality of audiences built on the Akkio + Affinity Solutions (AFS) data platform. The framework provides a standardized, repeatable methodology for evaluating how well any audience — whether a known-shopper seed, a lookalike expansion, or a propensity-scored cohort — predicts real consumer purchase behavior during a future holdout period.

The goal is to provide **apples-to-apples comparisons** across audience construction methods, enabling data-driven decisions about which approach delivers the highest-quality audiences for activation.

> **Note on platform capabilities:** The audiences delivered in this POC were built using Akkio's lookalike modeling and a deterministic SQL-based similarity scoring approach. Akkio's platform also includes a **dedicated propensity modeling engine** and a **built-in RFM feature engine** that were not utilized in this engagement. These capabilities — purpose-built for purchase propensity prediction with automated feature engineering, model training, and temporal holdout management — can further improve audience quality, particularly for niche brands where general-purpose similarity scoring has structural limitations. Integrating these tools into the workflow is a natural next step beyond this POC.

Key Concepts

What We Measure

Every audience is evaluated against the same set of metrics, calculated over a **holdout window** — a time period that was deliberately excluded when the audience was built. This ensures we are measuring true predictive power, not just historical behavior.

Metric	Formula	What It Tells You
Total Audience IDs	Count of distinct IDs in the audience	Audience reach / size
Active Matched IDs	Audience IDs with any transaction in the holdout period	Observable panel coverage within the audience
Brand Shoppers	Audience IDs with at least one brand transaction in holdout	How many audience members actually purchased
Brand Transactions	Total brand transactions from audience members in holdout	Volume of purchase activity
Brand Spend	Total dollar amount of brand transactions in holdout	Revenue impact
Shop Rate	Brand Shoppers / Active Matched IDs	% of active audience members who purchased the brand

Metric	Formula	What It Tells You
Spend Rate	Brand Spend / Active Matched IDs	Average brand spend per active audience member
Avg Ticket	Brand Spend / Brand Transactions	Average purchase size
Avg Transactions per Shopper	Brand Transactions / Brand Shoppers	Purchase frequency among converters

> **Why Active Matched IDs is the denominator (not Total Audience IDs):**
> Not every individual in the panel will have observable transaction activity in any given month. Using Total Active Matched IDs — audience members who had at least one transaction of *any* kind during the holdout — ensures we measure audience quality against observable behavior, rather than penalizing for panel coverage gaps.

Seed vs. Lookalike — Why Both Matter

Audience Type	What It Measures	Expected Behavior
Seed (Known Shoppers)	Retention / repeat purchase — "Do existing buyers keep buying?"	Higher shop rate (selecting on known buyers)
Lookalike / Propensity	Acquisition / prediction — "Can we find NEW brand shoppers before they buy?"	Lower shop rate, but demonstrates true predictive lift

A **seed audience** of known brand shoppers validated in a holdout window measures **retention** — these individuals already purchased, so the shop rate will be naturally high. This is useful for understanding audience stickiness but does not demonstrate predictive modeling power.

A **lookalike audience** or **propensity-scored cohort** of individuals who have *not* been observed purchasing the brand measures **acquisition** — the ability to identify future brand shoppers before they convert. This is the more meaningful test of audience quality and the direct comparison between modeling approaches.

Validation Methodology

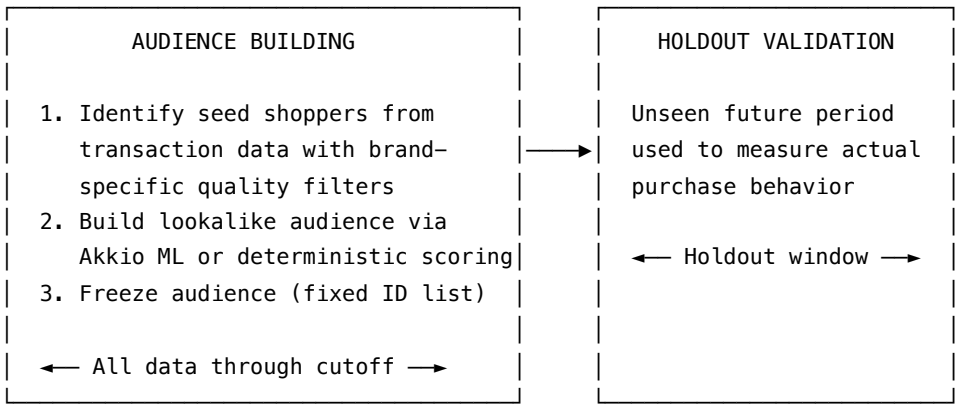
Timeline Split

The validation uses a temporal holdout design to measure audience quality against future purchase behavior:

Period	Role
All data through the build cutoff date	Used for audience construction (seed identification, feature engineering, model training)
Holdout window (post-cutoff)	Used **exclusively** for measuring audience performance — never seen during audience building

For the pre-delivery validation cycle, the holdout window is **September 1, 2025 through September 30, 2025**. AFS will conduct a definitive out-of-sample validation against the **October – December 2025** transaction window.

How Validation Works



Step-by-step process:

1. **Identify seed shoppers** — match brand keywords against transaction data with quality filters (minimum transaction thresholds, date bounds, holiday exclusions where applicable)
2. **Build the lookalike audience** — seed is used to generate a scored, ranked audience via Akkio's ML platform (LAL) or deterministic SQL similarity scoring across RFM + demographic + interest features
3. **Freeze the audience** — the list of IDs is fixed before looking at holdout data
4. **Join audience IDs** against all transactions in the holdout window
5. **Identify active members** — audience IDs that had *any* transaction (any brand) during the holdout
6. **Identify brand converters** — audience IDs that had at least one brand-specific transaction during the holdout
7. **Compute metrics** — Shop Rate, Spend Rate, Avg Ticket, Avg Transactions per Shopper, and Lift vs. general population baseline

Brand Matching Logic

Brand transactions are identified by matching keywords (case-insensitive) against three columns in the transaction data:

- `BRAND_NAME`
- `STORE_NAME`
- `MERCHANT_DESCRIPTION`

For example, to identify ActBlue transactions, the keyword `ACTBLUE` is matched against all three columns. Multiple keywords per brand are supported and OR'd together.

Delivered Audiences — Validation & Results

Overview

Seven final production audiences have been built, validated, and **delivered to the AFS S3 bucket** for activation. These audiences span four brands/segments — **ActBlue** (two construction methods), **Ross** (two construction methods), **BetMGM** (two construction methods), and **Holiday Discount Department Store Shoppers**.

AFS will conduct their own out-of-sample validation against the **October – December 2025** transaction window, a period not available during audience construction.

Final Audience Inventory

#	Audience Name	Type	Size	Brand(s)
1	ActBlue Contributor Seed Lookalike - Final	Akkio LAL	874,141	ActBlue
2	ActBlue Deterministic Lookalike Audience - Final	Deterministic SQL Scoring	500,000	ActBlue
3	Ross Seed Audience Lookalike - Final	Akkio LAL	3,075,285	Ross
4	Ross Deterministic Lookalike Audience - Final	Deterministic SQL Scoring	500,000	Ross
5	BetMGM Seed Lookalike - Final	Akkio LAL	340,999	BetMGM
6	BetMGM Deterministic Lookalike Audience - Final	Deterministic SQL Scoring	100,000	BetMGM
7	Holiday Discount Dept Store Shoppers Seed Lookalike - Final	Akkio LAL	7,022,972	Ross, TJMaxx, Marshalls, Burlington, Nordstrom Rack, Target, Walmart, Kohl's, Macy's, JCPenney

Holdout Validation Results

The following results were generated by validating each audience against the **September 2025** transaction holdout window (Sep 1 – Sep 30). Metrics are computed using the same standardized methodology described earlier in this document.

Audience Metrics

Metric	ActBlue LAL	ActBlue Deterministic LAL	Ross Seed LAL	Ross Deterministic LAL	BetMGM LAL	BetMGM Deterministic LAL	Holiday Discount Store LAL
Total Audience IDs	874,141	500,000	3,075,285	500,000	340,999	100,000	7,022,972
Active Matched IDs	228,581	497,399	2,919,651	499,144	79,769	98,628	6,576,908
Brand Shoppers	63,771	6,948	1,029,549	82,862	25,336	1,485	5,343,089
Brand Transactions	242,142	22,051	1,716,801	130,840	371,524	13,275	28,063,998

Metric	ActBlue LAL	ActBlue Deterministic LAL	Ross Seed LAL	Ross Deterministic LAL	BetMGM LAL	BetMGM Deterministic LAL	Holiday Discount Store LAL
Brand Spend	$5.5M	$454K	$117.5M	$7.5M	$37.0M	$874K	$1.63B
Shop Rate	**27.90%**	**1.40%**	**35.26%**	**16.60%**	**31.76%**	**1.51%**	**81.24%**
Spend Rate	**$24.18**	**$0.91**	**$40.25**	**$15.08**	**$464.34**	**$8.86**	**$248.25**
Avg Ticket	$22.83	$20.57	$68.44	$57.55	$99.70	$65.86	$58.18
Avg Trans / Shopper	3.80	3.17	1.67	1.58	14.66	8.94	5.25

Baseline Comparison & Lift

Metric	ActBlue LAL	ActBlue Deterministic LAL	Ross Seed LAL	Ross Deterministic LAL	BetMGM LAL	BetMGM Deterministic LAL	Holiday Discount Store LAL
Baseline Active IDs	43,114,357	43,114,357	43,114,357	43,114,357	43,114,357	43,114,357	43,114,357
Baseline Brand Shoppers	121,775	121,775	2,799,977	2,799,977	57,283	57,283	20,446,678
Baseline Shop Rate	0.28%	0.28%	6.49%	6.49%	0.13%	0.13%	47.42%
Shop Rate Lift	**98.8x**	**4.9x**	**5.4x**	**2.6x**	**239.1x**	**11.3x**	**1.7x**
Spend Rate Lift	**101.6x**	**3.8x**	**6.1x**	**2.3x**	**291.5x**	**5.6x**	**2.3x**

Key Observations

ActBlue Akkio LAL (98.8x lift) — The audience concentrates 63,771 political donors out of an 874K audience, achieving a 27.9% shop rate versus a 0.28% general population base rate. The high lift reflects the niche, highly distinctive nature of political donation behavior — donors differ sharply from the general population on both transaction patterns and demographic/interest attributes, making them highly identifiable via lookalike modeling. Average transactions per shopper (3.8) confirms the recurring donation pattern that makes this segment particularly valuable for media targeting.

ActBlue Deterministic LAL (4.9x lift) — The deterministic scoring approach applied to ActBlue delivers a 1.40% shop rate in a 500K audience — 4.9x above the general population baseline. Like BetMGM, the gap between the Akkio LAL (98.8x) and the deterministic approach (4.9x) reflects the challenge of identifying niche cause/donation behavior through general-purpose RFM + demographic similarity. Political donors don't look distinct in their overall shopping patterns — what distinguishes them is the

specific act of donating. That said, the 6,948 donors the deterministic approach identifies are genuine: they average 3.17 transactions per shopper and a $20.57 average ticket, consistent with ActBlue's recurring small-dollar donation pattern. The Akkio LAL remains the primary ActBlue audience; the deterministic variant provides an additional transparent, reproducible option.

Ross Seed LAL (5.4x lift) — The largest audience at 3.1M IDs, with over 1 million brand shoppers in the holdout month. The 35.3% shop rate at 5.4x lift demonstrates that Ross shopping behavior is highly habitual and the lookalike model effectively identifies repeat shoppers at scale. The spend rate ($40.25 per active member) represents meaningful media-targetable value.

Ross Deterministic LAL (2.6x lift) — The deterministic SQL-based scoring approach delivers a 16.6% shop rate in a tightly-sized 500K audience. While the lift (2.6x) is more conservative than the Akkio LAL, this audience was built using a fully transparent, reproducible scoring methodology (Gaussian similarity across all RFM + demographic features) with no black-box model. The two Ross audiences offer a precision/reach trade-off: the Deterministic LAL is the higher-precision, smaller audience; the Akkio LAL is the broader-reach, higher-volume option.

BetMGM Akkio LAL (239.1x lift) — The standout audience by lift, driven by sports betting's extremely low general population base rate (0.13%). The 31.8% shop rate means nearly 1 in 3 active audience members placed a bet in the holdout month. Average transactions per shopper (14.66) and average ticket ($99.70) reflect the high-frequency, high-value nature of sports betting — this audience has substantial economic density for activation.

BetMGM Deterministic LAL (11.3x lift) — The deterministic scoring approach applied to BetMGM delivers a 1.51% shop rate in a focused 100K audience — still 11.3x above the general population baseline. While the shop rate is significantly lower than the Akkio LAL (1.51% vs. 31.8%), this reflects the fundamental challenge of identifying a niche behavior (sports betting) through general RFM + demographic similarity alone. Notably, the bettors identified by the deterministic approach still show high engagement: 8.94 transactions per shopper and a $65.86 average ticket, confirming these are real, active bettors — just fewer of them. The two BetMGM audiences mirror the Ross pattern: the Akkio LAL is the high-reach option; the Deterministic LAL offers a smaller, transparent alternative with meaningful lift.

Holiday Discount Department Store Shoppers (1.7x shop lift, 2.3x spend lift) — The broadest audience at 7M IDs, covering a multi-brand retail category (Ross, TJMaxx, Marshalls, Burlington, Nordstrom Rack, Target, Walmart, Kohl's, Macy's, JCPenney). The 81.2% shop rate reflects the ubiquity of discount retail shopping — nearly half the general population already shops these stores in any given month (47.4% baseline). The 2.3x spend lift indicates the audience captures higher-spending shoppers within this high-base-rate category. This is a reach-oriented audience designed for broad holiday retail activation.

Comparison to AFS Purchase Propensity Modeling Approach

AFS employs a structured ML propensity modeling pipeline with a strict temporal holdout design:

Step	AFS Propensity Approach	Akkio Lookalike Approach
Label creation	Binary classification — shoppers (positive) vs. non-shoppers (negative) identified from Sept 2024 – Sept 2025	Seed identification — brand shoppers identified from transaction data with quality filters
Feature engineering	Historical features from Sept 2023 – Sept 2024 (1-year temporal gap between features and labels)	Pre-materialized RFM features (5 time windows) + 50+ demographic/interest/propensity attributes
Modeling	Two separate models: demographics-only and RFM-only; evaluated via AUC, Lift, and top-decile precision	Gaussian similarity scoring weighted by seed-vs-population divergence (deterministic), or Akkio's ML platform (LAL)

Step	AFS Propensity Approach	Akkio Lookalike Approach
Output	Probability-ranked audience segments (top 1%, 5%, 10%)	Similarity-ranked audience with configurable size cutoff
Temporal separation	1-year gap between feature window and label window	Features and seeds use overlapping data (see note below)

The two approaches are **complementary**:

- **AFS's propensity model** uses a rigorous 1-year temporal separation between features and labels, which produces unbiased estimates of predictive power. The separate demographics-only and RFM-only models allow attribution of which feature classes drive the most signal per brand.
- **The Akkio/deterministic approach** incorporates a broader feature set in a single unified score (RFM across 5 windows + all demographics + interests + propensities), which can capture complex multi-factor patterns. The trade-off is that the current validation cycle has some temporal overlap between features/seeds and the holdout window (see below).

For the Oct-Dec validation window, both approaches will be evaluated under identical conditions — a true out-of-sample test where neither approach had access to the holdout data. This will provide a clean, apples-to-apples comparison of audience quality.

Validation Caveats & Oct-Dec Expectations

Temporal overlap in current validation: The September holdout validation presented above has a known limitation — the RFM features and seed identification both use data through end of September 2025, which overlaps with the September holdout window. This means:

1. Seed members identified from September transactions are captured in the lookalike → they will naturally appear as brand shoppers in the September holdout (circular)
2. RFM features (especially 1-month and 3-month windows) incorporate September behavior, giving the scoring model information about the holdout period

As a result, the shop rates and lifts above are **optimistic** relative to a clean out-of-sample test. They demonstrate that the audiences are correctly constructed and enriched for the target behaviors, but the exact magnitudes should be interpreted as upper bounds.

Why we remain confident for Oct-Dec:

1. **Behavioral persistence** — The brands in this portfolio exhibit highly habitual consumer behavior. Political donors (ActBlue) give repeatedly. Retail shoppers (Ross) at 3+ transactions per year are habitual visitors. Sports bettors (BetMGM) at 14.7 transactions/month are deeply engaged. These patterns persist month-over-month and are not artifacts of a single month's data.
2. **Lift headroom** — Even with a significant discount for temporal overlap, the lift numbers have substantial margin. ActBlue at 99x and BetMGM at 239x could lose 50-70% of their lift and still represent exceptional audiences. Ross at 5.4x could halve and remain a strong 2-3x performer.
3. **Behavioral signatures pass sanity checks** — Average ticket values are consistent with brand economics (Ross ~ $68, $ActBlue$ 23, BetMGM ~$100). Transactions per shopper reflect real behavioral patterns (3.8 recurring donations vs. 1.6 retail visits vs. 14.7 betting sessions). These would not be coherent if the audiences were purely artifacts of leakage.
4. **AFS's Oct-Dec validation is the definitive test** — The October through December window was not available during any stage of audience construction (seed identification, feature engineering, or scoring). This will provide fully unbiased metrics

against which to measure audience quality and compare approaches.

Expected Oct-Dec performance ranges:

Audience	Sept Shop Rate	Expected Oct-Dec Range	Rationale
ActBlue LAL	27.9%	12-20%	Recurring donors persist; some Sept-only donors drop off
ActBlue Deterministic LAL	1.40%	0.5-1.0%	Niche cause behavior harder to capture deterministically; recurring donors provide baseline lift
Ross Seed LAL	35.3%	15-25%	Habitual retail; holiday season may boost or maintain
Ross Deterministic LAL	16.6%	8-14%	More conservative baseline; most honest current signal
BetMGM LAL	31.8%	15-25%	NFL season Oct-Dec is peak betting; strong tailwind
BetMGM Deterministic LAL	1.51%	0.5-1.2%	Niche behavior harder to capture deterministically; still meaningful lift expected
Holiday Discount Store LAL	81.2%	65-80%	Holiday season is the target period; baseline also rises

Delivery Status

All seven final audiences have been delivered to the AFS S3 bucket for activation and validation. Each audience file contains the audience member AFS_IDs as scored and ranked by the respective modeling approach.

How to Interpret These Results

Shop Rate Is the Primary Quality Signal

Shop Rate answers the core question: "*What percentage of active audience members actually purchased from the brand during the holdout window?*"

- **Higher shop rate = better audience quality** — the audience is more concentrated with likely buyers
- **Seed audiences will always have higher shop rates** — they are built from known buyers, so holdout shop rate measures retention, not prediction
- **Lookalike / propensity audiences are the true test** — their shop rate reflects the model's ability to identify *new* buyers

Spend Rate Captures Dollar-Value Impact

Spend Rate combines shop rate with purchase value to answer: "*How much brand revenue does each active audience member represent?*"

This is the most actionable metric for media planning — it directly translates audience quality to economic value.

Lift Over Baseline

To contextualize these results, compare against a **random baseline** — the shop rate you would observe by taking a random sample of the same size from the active population. Any audience's shop rate divided by the baseline shop rate gives you **lift**:

```
Lift = Audience Shop Rate / Random Baseline Shop Rate
```

A lift of 2.0x means the audience is twice as likely to contain brand shoppers as a random sample. Higher lift = better targeting precision.

Summary

The audience validation framework provides a rigorous, transparent method for measuring audience quality:

1. **Temporal holdout design** ensures metrics reflect true predictive power by evaluating audiences against unseen future transaction data
2. **Standardized metrics** (Shop Rate, Spend Rate, Avg Ticket, Lift) enable direct comparison across audiences, brands, and modeling approaches
3. **Active Matched IDs denominator** ensures fair comparison regardless of panel coverage differences
4. **Baseline lift comparison** quantifies how much better each audience performs relative to a random sample of the general population
5. **Cross-methodology comparison** enables evaluation of Akkio's ML-based lookalike modeling alongside deterministic similarity scoring and AFS's propensity modeling approach

All seven delivered audiences demonstrate meaningful lift over the general population baseline, with the strongest results for niche/cause brands (ActBlue, BetMGM) where the target behavior is highly distinctive. AFS's upcoming October – December validation will provide the definitive out-of-sample test of audience quality.

Appendix: Lookalike Audience Design Rationale

The following sections provide detailed design rationale and reference SQL for the deterministic lookalike methodology. The compact rules for query generation live in `affinity_client_context.md` ; this appendix explains the *why* behind those rules.

Why Scoring Beats Thresholds

Binary pass/fail thresholds on a limited feature set (e.g., RFM-only) have three structural problems:

1. **No gradient** — A prospect who barely misses one threshold is treated the same as one who misses everything. There is no "how similar" signal.
2. **Limited signal** — Using only RFM for matching discards the strongest predictors for many brands. For niche/cause brands (e.g., ActBlue), demographics and interests are often more discriminative than general transaction behavior. For retail brands, channel and category affinity may matter as much as spend.
3. **No ranking** — Everyone who clears the thresholds is "equally good," with no way to prioritize the best prospects or control audience size precisely.

A scoring approach solves all three: every feature contributes a weighted signal, prospects are ranked by total similarity, and audience size is controlled by taking the top N. Seed members naturally score high (they match the seed profile by definition) without being force-included, producing honest validation metrics.

Seed Member Handling

Seed members are included in the scored population and are NOT force-excluded. They naturally rank near the top because they match the seed profile. This mirrors how propensity models work:

- A propensity model scores everyone; known buyers score well organically
- Force-excluding seed creates an artificial gap and can bias the LAL toward lower-quality prospects
- Force-including seed (as a union) inflates validation metrics without reflecting LAL quality

When validating, segment results by seed vs. non-seed to measure incremental LAL lift separately. The delivered audience should contain seed members naturally ranked by score.

Precision vs. Reach Tuning

Audience size (`<audience_size>`) directly controls the precision/reach trade-off:

Goal	Audience Size	Expected Outcome
High Precision	Small (top 50K–100K)	Highest-scoring prospects; strong similarity, highest expected conversion
Balanced	Medium (top 250K–500K)	Good similarity with broader reach; moderate conversion
High Reach	Large (top 1M+)	Wider net; scores taper, conversion approaches baseline at the margin

The score distribution itself is informative: a steep drop-off means the seed is highly distinctive and a tight audience is appropriate. A gradual decline means the seed blends with the population and larger audiences are needed.

Small Seed Safeguards — Detailed Explanation

Niche brands (political donations, sports betting, luxury goods) often produce very small seeds — sometimes fewer than 10 people. The scoring methodology handles this through two critical floors:

Bandwidth Floor: `GREATEST(COALESCE(seed_std, 0), 0.5 * pop_std)`

- Seed size = 1: `STDDEV()` returns NULL → falls back to `0.5 * pop_std`
- Seed size = 2–10 with similar values: `STDDEV()` near-zero → Gaussian becomes needle-thin → falls back to `0.5 * pop_std`
- Normal seed variance: `seed_std > 0.5 * pop_std` → uses `seed_std` (no change)
- The 0.5 multiplier means the Gaussian is twice as selective as the population spread — still providing meaningful differentiation while preventing collapse

Importance Floor: `GREATEST(importance, 0.1)`

- Very large/common seeds (e.g., "all holiday shoppers"): Seed ≈ population → importance ≈ 0 → floor ensures Gaussian still differentiates by proximity to seed mean

Seed Count Diagnostic: Always output seed count. If seed is empty (0 members), return informative message rather than misleading all-zero audience.

Key Design Points

- **Pre-materialized RFM via `RFM_FEATURES`:** Scoring reads pre-computed features — no transaction-level aggregation at scoring time, eliminating the biggest performance bottleneck.
- **Single CTE chain:** The entire query is one `CREATE TABLE AS WITH ... SELECT ...` statement. No temp tables. Since `RFM_FEATURES` eliminates the heavy scan, a CTE chain executes efficiently.
- **Pre-computed categorical importance:** Each `IMP_*` CTE computes the scalar importance once. The scoring SELECT references these scalars instead of re-evaluating correlated subqueries per row.
- **No expression duplication:** `NUMERIC_SIMILARITY_SCORE` and `CATEGORICAL_SIMILARITY_SCORE` are computed once in the `SCORED` CTE, then summed in the final SELECT.
- **Gaussian similarity** (`EXP(-0.5 * z²)`) produces a smooth 0-to-1 score per feature. A prospect matching the seed mean exactly scores 1.0; distant values taper toward 0.
- **Automatic feature weighting** via `importance = |seed_mean - pop_mean| / pop_stddev` — no manual weight tuning needed.
- **All features contribute** — RFM (all 5 windows), demographics, interests, propensities. The weighting ensures discriminative features matter most.

Reference SQL — Full Lookalike Scoring Query

The following is the complete SQL pattern for a lookalike audience build. The compact version in `affinity_client_context.md` describes the CTE structure; this is the fully expanded reference.

```
CREATE TABLE <output_table> AS
WITH
-- Phase 1: Seed identification & population features
SEED_IDS AS (
  SELECT DISTINCT AKKIO_ID
  FROM FACT_TRANSACTION_ENRICHED
  WHERE (<brand_filter>)
    AND TRANS_DATE >= DATEADD(MONTH, -12, '<ref_date>'::DATE)
    AND TRANS_DATE < '<cutoff_day_after>'
),

POP_FEATURES AS (
  SELECT
    r.AKKIO_ID,
    r.days_since_last_txn,
    r.tot_trans_12mo, r.tot_spend_12mo,
    r.tot_online_trans_12mo, r.tot_online_spend_12mo,
    r.avg_days_btwn_trans_12mo, r.brand_diversity_12mo, r.online_ratio_12mo,
    r.tot_trans_9mo, r.tot_spend_9mo, r.tot_online_trans_9mo, r.tot_online_spend_9mo,
    r.avg_days_btwn_trans_9mo, r.brand_diversity_9mo,
    r.tot_trans_6mo, r.tot_spend_6mo, r.tot_online_trans_6mo, r.tot_online_spend_6mo,
    r.avg_days_btwn_trans_6mo, r.brand_diversity_6mo,
    r.tot_trans_3mo, r.tot_spend_3mo, r.tot_online_trans_3mo, r.tot_online_spend_3mo,
    r.avg_days_btwn_trans_3mo, r.brand_diversity_3mo,
    r.tot_trans_1mo, r.tot_spend_1mo, r.tot_online_trans_1mo, r.tot_online_spend_1mo,
    r.avg_days_btwn_trans_1mo, r.brand_diversity_1mo,
    CASE WHEN s.AKKIO_ID IS NOT NULL THEN 1 ELSE 0 END AS IS_SEED,
    d.GENDER, d.STATE, d.POLITICS, d.INCOME_BUCKET, d.EDUCATION_LEVEL,
    d.ETHNICITY, d.AGE, d.MARITAL_STATUS, d.HOMEOWNER_STATUS,
    d.NET_WORTH_BUCKET, d.OCCUPATION
    -- ... include ALL demographic, interest, and propensity fields from V_AKKIO_ATTRIBUTES_LATEST
  FROM RFM_FEATURES r
  LEFT JOIN SEED_IDS s ON r.AKKIO_ID = s.AKKIO_ID
  LEFT JOIN V_AKKIO_ATTRIBUTES_LATEST d ON r.AKKIO_ID = d.AKKIO_ID
),

-- Phase 2: Statistics & scoring
SEED_NUMERIC_STATS AS (
  SELECT
    AVG(days_since_last_txn) AS seed_mean_recency,
    STDDEV(days_since_last_txn) AS seed_std_recency,
    AVG(tot_trans_12mo) AS seed_mean_freq,
    STDDEV(tot_trans_12mo) AS seed_std_freq,
    AVG(tot_spend_12mo) AS seed_mean_spend,
    STDDEV(tot_spend_12mo) AS seed_std_spend,
    AVG(tot_trans_3mo) AS seed_mean_freq_3mo,
    STDDEV(tot_trans_3mo) AS seed_std_freq_3mo,
    AVG(online_ratio_12mo) AS seed_mean_online_ratio,
    STDDEV(online_ratio_12mo) AS seed_std_online_ratio,
    AVG(brand_diversity_12mo) AS seed_mean_brand_div,
    STDDEV(brand_diversity_12mo) AS seed_std_brand_div,
    AVG(avg_days_btwn_trans_12mo) AS seed_mean_cadence,
    STDDEV(avg_days_btwn_trans_12mo) AS seed_std_cadence,
    AVG(CAST(AGE AS FLOAT)) AS seed_mean_age,
```

```
    STDDEV(CAST(AGE AS FLOAT)) AS seed_std_age
    -- ... repeat for ALL numeric features
  FROM POP_FEATURES
  WHERE IS_SEED = 1
),

POP_NUMERIC_STATS AS (
  SELECT
    AVG(days_since_last_txn) AS pop_mean_recency,
    STDDEV(days_since_last_txn) AS pop_std_recency,
    AVG(tot_trans_12mo) AS pop_mean_freq,
    STDDEV(tot_trans_12mo) AS pop_std_freq,
    AVG(tot_spend_12mo) AS pop_mean_spend,
    STDDEV(tot_spend_12mo) AS pop_std_spend,
    AVG(tot_trans_3mo) AS pop_mean_freq_3mo,
    STDDEV(tot_trans_3mo) AS pop_std_freq_3mo,
    AVG(online_ratio_12mo) AS pop_mean_online_ratio,
    STDDEV(online_ratio_12mo) AS pop_std_online_ratio,
    AVG(brand_diversity_12mo) AS pop_mean_brand_div,
    STDDEV(brand_diversity_12mo) AS pop_std_brand_div,
    AVG(avg_days_btwn_trans_12mo) AS pop_mean_cadence,
    STDDEV(avg_days_btwn_trans_12mo) AS pop_std_cadence,
    AVG(CAST(AGE AS FLOAT)) AS pop_mean_age,
    STDDEV(CAST(AGE AS FLOAT)) AS pop_std_age
    -- ... same fields as SEED_NUMERIC_STATS
  FROM POP_FEATURES
),

-- Categorical distribution + importance (repeat pattern for every categorical field)
SEED_CAT_GENDER AS (
  SELECT GENDER AS cat_value, COUNT(*)::FLOAT / SUM(COUNT(*)) OVER () AS seed_share
  FROM POP_FEATURES WHERE IS_SEED = 1 AND GENDER IS NOT NULL GROUP BY GENDER
),
POP_CAT_GENDER AS (
  SELECT GENDER AS cat_value, COUNT(*)::FLOAT / SUM(COUNT(*)) OVER () AS pop_share
  FROM POP_FEATURES WHERE GENDER IS NOT NULL GROUP BY GENDER
),
IMP_GENDER AS (
  SELECT GREATEST(MAX(sc.seed_share / NULLIF(pc.pop_share, 0)), 0.1) AS importance
  FROM SEED_CAT_GENDER sc JOIN POP_CAT_GENDER pc ON sc.cat_value = pc.cat_value
),
-- ... repeat SEED_CAT_, POP_CAT_, IMP_ for: STATE, POLITICS, INCOME_BUCKET,
--     EDUCATION_LEVEL, ETHNICITY, MARITAL_STATUS, HOMEOWNER_STATUS,
--     NET_WORTH_BUCKET, OCCUPATION, and ALL other categorical fields

SCORED AS (
  SELECT
    P.AKKIO_ID,
    P.IS_SEED,

    -- Numeric: EXP(-0.5 * ((val - seed_mean) / bandwidth)^2) * importance
    EXP(-0.5 * POW((P.days_since_last_txn - S.seed_mean_recency)
        / GREATEST(COALESCE(S.seed_std_recency, 0), 0.5 * POP.pop_std_recency), 2))
      * GREATEST(ABS(S.seed_mean_recency - POP.pop_mean_recency) / NULLIF(POP.pop_std_recency, 0), 0.1)
```

```
    + EXP(-0.5 * POW((P.tot_trans_12mo - S.seed_mean_freq)
        / GREATEST(COALESCE(S.seed_std_freq, 0), 0.5 * POP.pop_std_freq), 2))
      * GREATEST(ABS(S.seed_mean_freq - POP.pop_mean_freq) / NULLIF(POP.pop_std_freq, 0), 0.1)
    + EXP(-0.5 * POW((P.tot_spend_12mo - S.seed_mean_spend)
        / GREATEST(COALESCE(S.seed_std_spend, 0), 0.5 * POP.pop_std_spend), 2))
      * GREATEST(ABS(S.seed_mean_spend - POP.pop_mean_spend) / NULLIF(POP.pop_std_spend, 0), 0.1)
    -- + ... repeat for ALL other numeric features
      AS NUMERIC_SIMILARITY_SCORE,

    -- Categorical: seed_share * scalar importance from IMP_ CTEs
    COALESCE(SG.seed_share, 0) * (SELECT importance FROM IMP_GENDER)
    + COALESCE(SS.seed_share, 0) * (SELECT importance FROM IMP_STATE)
    -- + ... repeat for ALL other categorical fields
      AS CATEGORICAL_SIMILARITY_SCORE

  FROM POP_FEATURES P
  CROSS JOIN SEED_NUMERIC_STATS S
  CROSS JOIN POP_NUMERIC_STATS POP
  LEFT JOIN SEED_CAT_GENDER SG ON P.GENDER = SG.cat_value
  LEFT JOIN SEED_CAT_STATE SS ON P.STATE = SS.cat_value
  -- ... LEFT JOIN for each categorical seed distribution CTE
)

-- Phase 3: Rank and extract
SELECT
  AKKIO_ID, IS_SEED,
  NUMERIC_SIMILARITY_SCORE, CATEGORICAL_SIMILARITY_SCORE,
  (NUMERIC_SIMILARITY_SCORE + CATEGORICAL_SIMILARITY_SCORE) AS SIMILARITY_SCORE
FROM SCORED
ORDER BY SIMILARITY_SCORE DESC
LIMIT <audience_size>;
```