Problem Statement — Validate & Stress-Test LP Score v4

Background

You are given a wallet-level dataset of **LP (liquidity provider) scoring**. Each row is a wallet with:

- Top-level fields: `_id`, `wallet_id`, **`aggregated_lp_score`**, and three category fields:
    - `lp_category_breakdown.stable-stable`
    - `lp_category_breakdown.stable-volatile`
    - `lp_category_breakdown.volatile-volatile`
- Up to **13 per-wallet LP slots** (`lp_scores[0..12]`), each with pool metadata and metrics:
    - `pool_id`, `pool_name`, `fee_tier`, `token_symbols[0/1]`, `tvl`, `last_tx_timestamp`, `timestamp`
    - Activity & behavior: `num_deposits`, `num_withdrawals`, `avg/min/max_holding_days`, `liquidity_percent_remaining`, `retained_liquidity`, `lp_volatility_stddev`, `dust_deposit_count`, `dust_deposit_volume`, `total_deposit_all_time`, `total_withdraw_all_time`
    - Scoring: `score_breakdown.*` (components below) and **`total_score`** per pool
      Components include: `deposit_volume_score`, `withdraw_volume_score`, `deposit_frequency_score`, `avg_holding_time_score`, `liquidity_retention_score`, `lp_volatility_score`, `time_score`, and `score_breakdown.total_score` (matches `total_score`).

Scale snapshot (for orientation, not targets):

- Rows: **44,975**; Columns: **396**
- `aggregated_lp_score`: min **45.0**, median **296.65**, 90th **~498.95**, max **~937.25**
- Pools per wallet: median **1**, mean **~1.25**, max **13**

- Category breakdown fields are **non-negative, wide-range** integers; they **do not always sum** to `aggregated_lp_score` (**≈41%** exact matches observed).
- Sum of per-pool `total_score` **does not always equal** `aggregated_lp_score` (**≈41%** matches).
- Aggregate **deposit/withdraw volumes** have **weak linear correlation** with `aggregated_lp_score` (**~0.03**), implying non-volume factors matter (frequency, retention, volatility, time).

Objective

Validate, explain, and pressure-test how `aggregated_lp_score` is derived from per-pool behavior and the score components. Identify **anomalies, edge cases, and inconsistencies**—especially cases where behavior suggests a higher (or lower) score than assigned.

Key Questions to Answer

1. **Construction validity**
    - What is the precise relationship between `aggregated_lp_score` and the set of per-pool `total_score` values? Is there a normalization/weighting/attenuation step? Why do ~40% of rows match exactly but many don't?
    - Do the three `lp_category_breakdown.*` fields represent contributory sub-scores or something else? Should they sum to `aggregated_lp_score`?
2. **Behavior–score alignment**
    - Are **higher deposit volumes** or **more consistent liquidity retention** reflected in higher scores (monotonicity)? Quantify where this fails (e.g., top-decile depositors with below-median scores).
    - Do wallets with **high `liquidity_percent_remaining` / `retained_liquidity`** consistently get higher **liquidity_retention_score** and higher `aggregated_lp_score`?
    - How do **`dust_deposit_count/volume`** and **`lp_volatility_stddev`** influence scores? Are there wallets penalized/boosted disproportionately?
3. **Component coherence**
    - Within a pool, does `score_breakdown.total_score` equal the sum (or weighted sum) of its components? Confirm for all indices `[0..12]`.

    - Across pools, are component effects consistent (e.g., same deposit frequency → similar `deposit_frequency_score` scale)?

4. **Anomalies & outliers**
    - Identify wallets with **extreme mismatches** (e.g., high `total_deposit_all_time` + high retention but low score; or low activity but high score).
    - Spot **temporal oddities** (e.g., `last_tx_timestamp` far in the past but score remains unusually high).
    - Find **data quality** issues (impossible negatives, duplicates across `wallet_id` + pool, empty token symbols, etc.).

Required Analyses

- **Reconstructability check:** Attempt to reproduce `aggregated_lp_score` (e.g., sum of per-pool `total_score` with/without weights; try simple normalizations). Document fits/misses.
- **Cohort studies:** Bucket wallets by deposit/withdraw deciles, holding-time bands, and retention percentiles; chart median/quantile **aggregated scores** per cohort; flag monotonicity breaks.
- **Component attribution:** For top pools per wallet, quantify contribution of each `score_breakdown.*` to `total_score`; rank which components drive high scores.
- **Outlier surfacing:** Use robust methods (IQR/MAD) on key ratios (e.g., score per $ deposited, score per deposit, score vs. retention) to list anomalies with wallet IDs and pool context.
- **Category breakdown audit:** Test whether `lp_category_breakdown.*` fields act as partials of `aggregated_lp_score` or independent diagnostics; document discrepancies.

Deliverables

1. **Short report (2–3 pages)** with:
    - Executive bullet points (top 10 facts/anomalies).
    - Clear tables/plots proving each claim.
    - A concise explanation of the **most plausible formula** (or why it cannot be uniquely inferred).
2. **Reproducible code** (notebook or script) to regenerate the findings.

3. **Anomalies CSV**: `wallet_id, pool_id(optional), reason, metric, value, threshold`.
4. **Assumptions & open questions** you need clarified to finalize a formula.

Constraints & Evaluation

- **No AI/online tools**; original work only.
- Judged on **originality, rigor, clarity, and reproducibility**.
- Every claim must be **data-verifiable** (we will rerun your code).