Professional Data Cleaning & Engineering Log

Project: AMFI Mutual Fund Master Dataset Optimization

Analyst Team: Ranajeet Roy (Lead), Abhijeet Kumar (Lead), Yatin Singh, Pratham Malhotra, Pushkar Jain, Ved Pawar

Date: February 18, 2026

Dataset Scale: Total AUM Verified at 8,271,110.904 Cr

1. Structural Transformations & Error Recovery

Phase I – Scheme_Category: Mapped granular scheme labels into four macro categories: Equity, Debt, Hybrid, and Other.

Phase II – Scheme_NAV_Name: Column dropped due to naming noise such as Regular, Direct, IDCW, and Retail variations.

Phase III – Artifact Cleanup: Removed columns created by failed text-to-columns parsing, including symbol fragments like 0, ., /, ---, and \.

2. Numerical Processing & Outlier Management

Phase IV – Scheme_Min_Amt: Extracted numeric values and converted unit-based strings into absolute integers.

Phase V – Missing Value Treatment: Replaced 'REFER SID' and nulls with the median.

Phase VI – Outlier Treatment: Removed statistical anomalies in NAV and Scheme_Min_Amt.

3. Temporal Feature Engineering

Phase VII – Fund_Age: Derived fund age in years using datetime differences.

4. Final Data Integrity Summary

Dimensionality: Removed Scheme_NAV_Name and parsing artifacts.

Distribution Health: Cleaned NAV and Scheme_Min_Amt of outliers.

Analytical Readiness: Dataset ready for AUM, category, and fund-age correlation analysis.