

Subject: Data Quality Issues, Key Trends & Next Steps – Action Needed

Hi [Product/Business Leader],

I've completed an initial assessment of the User, Transactions, and Products datasets, identifying key data quality issues affecting analysis. Additionally, I've analyzed user signups by generation, uncovering an interesting trend that may inform future strategy. Please see detailed analysis in attachment.

Would love to set up a quick sync to align on these next steps. Let me know a time that works for you.
Thanks!

Best,
Li Li

-----[Attachment]-----

1. Key Data Quality Issues

1.1 Missing User & Product Links

- 88% of transactions (44,000 records) contain USER_IDs that don't exist in the User table.
- 38.82% of transactions (19,412 records) have BARCODEs that don't match any Products.

Impact: Limits our ability to analyze customer behavior and product performance.

1.2 Data Gaps & Inconsistencies

- Missing Values: BIRTH_DATE, GENDER, and STATE in User; CATEGORY_4, BRAND, MANUFACTURER, and primary key BARCODE in Products; BARCODE in Transactions.
- Duplicate & Conflicting Entries: RECEIPT_ID alone is not a unique identifier—a single transaction can contain multiple products. Primary key should be (RECEIPT_ID, BARCODE) instead of RECEIPT_ID.

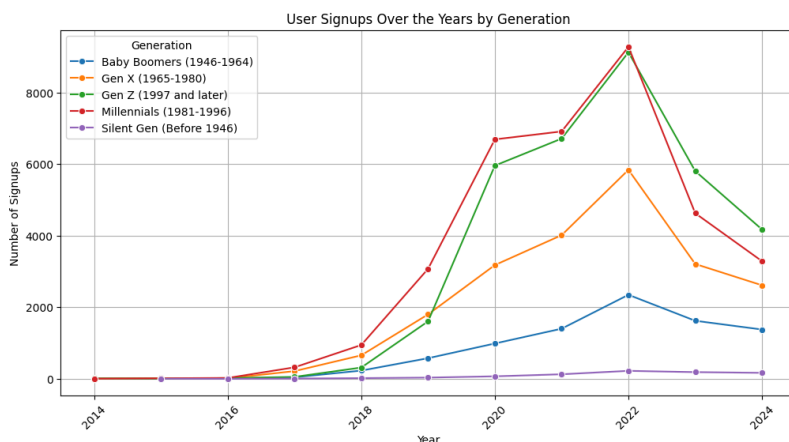
Impact: Could lead to incorrect deduplication and misrepresented sales data.


1.3 Unexpected Data Patterns


- Outliers: Some transactions have near-zero-unit prices or fractional FINAL_QUANTITY values (which may indicate weight-based pricing).
- Unrealistic User Age Entries: Ages range from 0 to 121, and placeholder birthdates.
- Time Zone Mismatch & Negative Time Differences:
 - SCAN_DATE, CREATED_DATE, and BIRTH_DATE in UTC, while PURCHASE_DATE lacks a time zone.
 - Even after standardizing timestamps, some receipts are scanned before the purchase date, suggesting system lags or data entry errors.

Impact: Raises concerns about transaction validity, pricing accuracy, and system issues.

2. Key Trend: User Signups by Generation



-  Peak Growth (2020-2022):
- Millennials (1981-1996) & Gen Z (1997+) drove the most signups, peaking in 2022.
 - Gen X & Baby Boomers showed steady but slower adoption.
 - Silent Gen adoption was minimal (low tech adoption).

 Post-2022 Decline:

- Likely due to market saturation, economic shifts, or reduced marketing effectiveness.

Next Steps:

- ✓ Focus on retention strategies for Millennials & Gen Z.
- ✓ Explore new acquisition tactics for Gen X & Baby Boomers.
- ✓ Analyze 2022 marketing campaigns to understand peak growth drivers.

3 . *Request for Action*

We need input from the Data Engineering team to:

- ✓ Investigate why so many transactions lack valid users and product links.
- ✓ Confirm (RECEIPT_ID, BARCODE) as the correct primary key for Transactions table.
- ✓ Investigate missing values across the Users, Products, and Transactions tables.
- ✓ Clarify if fractional and FINAL_QUANTITY and extremely low unit price are expected.
- ✓ Validate whether users' extreme ages and placeholder birth dates were self-reported, or system generated.
- ✓ Confirm whether all timestamps should be converted to Central Time (CST) and investigate potential system lags or data entry errors causing receipts to be scanned before the purchase date.