# Data Cleaning & Preparation Report – Users Table

Document Date: November 2025 (Based on data timestamps)

Dataset: Users Table

Goal: Achieve a business-ready, consistent, and optimized dataset for analytics and dashboard visualization.

## 1.0 Introduction

This document provides a comprehensive record of all data preparation and cleaning steps performed on the Users Table dataset. The goal was to ensure the dataset is business-ready, consistent, and optimized for analytics and dashboard visualization by addressing nulls, removing technical noise, and resolving duplicates.

## 2.0 Column Exclusion: Metadata & Technical Keys

Ten columns were removed as they were identified as purely technical keys, metadata, or unreadable system references that lacked direct analytical value for understanding user behavior.

### 2.1 Technical / Metadata Columns Removed

| **Column** | **Category** | **Rationale for Removal** | **Alternative for Analysis** |
| --- | --- | --- | --- |
| instance\_id | System ID | Internal database key; no business value. | Use client\_name/site\_name |
| generated\_for\_date | ETL Metadata | Snapshot timestamp, not user activity. | Use account\_created\_ts, last\_login\_ts |
| generated\_ts | ETL Metadata | Operational timestamp; irrelevant for analysis. | Not applicable |
| user\_type\_id | System Reference ID | Numeric ID requires lookup; not readable. | Use **user\_type** name |
| retailer\_id | System Reference ID | Numeric ID; lacks business meaning. | Use **retailer** name |
| location\_id | System Reference ID | Database key only; no geographical context. | Use country\_name, location\_state, location\_city |
| user\_rank\_id | System Reference ID | Maps to gamification rank; not immediately readable. | Use **user\_rank\_title** / **user\_rank\_number** |
| preferred\_language\_id | System Configuration | Not relevant for engagement metrics. | Not critical |
| origin\_auth\_source\_id | Authentication Detail | Technical only; does not affect behavior. | Not applicable |
| location\_metadata\_json | Raw JSON | Unstructured; complex parsing required. | Use structured location fields |

## 3.0 High-Null Column Removal

Seven columns exhibiting high sparsity ( $>75\%$ null values) were removed to maintain data integrity and focus on features with robust coverage.

### 3.1 Columns with $>75\%$ Null Values

| **Column** | **Null %** | **Reason for Removal** |
| --- | --- | --- |
| employee\_number | 85% | Sparse field; not critical for dashboard. |
| num\_sale\_submissions | 87% | Low data coverage; not primary engagement metric. |
| num\_approved\_sale\_submissions | 88% | Low data coverage; not primary engagement metric. |
| num\_rejected\_sale\_submissions | 87% | Low data coverage; not primary engagement metric. |
| num\_pending\_sale\_submissions | 86% | Low data coverage; not primary engagement metric. |
| referral\_qualified\_on\_ts | 80% | Sparse; low business relevance. |
| referred\_by\_retailer\_title | 75% | Limited coverage; irrelevant for analysis. |

## 4.0 Null Value Treatment

Missing values were systematically treated based on their data type and business meaning to prevent data loss and ensure accurate aggregation.

### 4.1 Numerical Columns – Filled with 0

| **Column** | **Reason for Filling with 0** |
| --- | --- |
| **num\_days\_online** | Null implies user **never accessed** the platform. |
| **has\_reverified** | Null implies **no re-verification activity** has occurred. |
| **total\_spendable\_points\_earned** | Null implies **zero points** have been earned. |

### 4.2 Categorical Columns – Filled with Meaningful Defaults

**Location and Referral fields** were explicitly marked with Unknown or No Referral to prevent nulls from skewing descriptive statistics.

| **Column** | **Filled With** | **Rationale** |
| --- | --- | --- |
| first\_verification\_method | **Not Verified** | Makes verification status explicit. |
| hierarchy\_title | **Not Specified** | Prevents nulls in role analysis. |
| **Location Fields** (10 total) | **Unknown** | Explicitly marks missing geography. |
| **Referral Fields** (5 total) | **No Referral** / **No Referrer** | Explicitly marks organic sign-ups. |
| last\_verified\_ts | **Never Verified** | Shows the user's current verification state. |
| verification\_expired\_ts | **Never Verified** | Shows the expiration state. |
| first\_li\_completion\_ts | **Never Completed** | Tracks learning completion status. |
| last\_li\_completion\_ts | **Never Completed** | Tracks learning completion status. |
| user\_rank\_number | **0** | Users without a rank start at 0. |
| user\_rank\_title | **No Rank** | Represents the starting gamification level. |

### 4.3 Meaningful Nulls Kept (Informative Timestamps)

The following timestamp columns were retained with their null values, as **null is an informative state** (e.g., null in account\_disabled\_ts means the account is *not* disabled):

* account\_deleted\_ts, account\_disabled\_ts
* first\_login\_ts, last\_login\_ts, last\_online\_ts, last\_app\_login\_ts

## 5.0 Data Type Conversions

Date/Time columns critical for timeline analysis were standardized to the datetime format.

| **Column** | **Original Type** | **Converted Type** | **Rationale** |
| --- | --- | --- | --- |
| last\_verified\_ts | object | **datetime** | Needed for timeline analysis. |
| account\_created\_ts | object | **datetime** | Ensures correct chronological operations. |
| first\_li\_completion\_ts | object | left as **text** | Retains "Never Completed" labels for dashboard categorization. |
| verification\_expired\_ts | object | left as **text** | Retains "Never Verified" labels. |
| last\_li\_completion\_ts | object | left as **text** | Retains "Never Completed" labels. |

## 6.0 Duplicate Handling

Duplicate records were resolved by identifying and retaining the latest unique record for each user\_id, ensuring the final dataset represents unique entities.

* **Original sample:** 100,000 rows
* **Duplicate user\_id records found:** 30,812
* **Final cleaned dataset (Unique Users):** **69,188**

## 7.0 Sample Limitation

* The analysis was performed on a sample of **100,000 rows** for efficient processing.
* **17+ high-null columns** were removed to focus analysis.
* **Future Pipeline Issues:** Known issues related to future timestamps (e.g., Nov 2025) from ETL processes may persist in the final data and require continuous monitoring.

## 8.0 Recommendations

Based on the preliminary data exploration and cleaning process, the following actions are recommended for the business and data pipeline improvement:

1. **Improve Verification Flow:** The large proportion of "Not Verified" and "Partially Verified" accounts (which were the most engaged) suggests a bottleneck in the current verification process.
2. **Launch Reactivation Campaigns:** Target the substantial group of inactive or disabled users identified in the time series analysis to boost overall active engagement.
3. **Strengthen Referral Program Adoption:** Given the dominance of "No Referral," the program's incentives or tracking mechanism requires urgent review and optimization.
4. **Redesign Gamification:** Address the high correlation between L1 completions, achievements, and badges, which indicates redundancy in the reward structure.
5. **Enrich Missing Geographic Data:** Develop a strategy to append or infer missing geographic data, which currently relies on the "Unknown" category for over 60% of records.
6. **Track Course Progression:** Implement tracking for intermediate learning progression events (e.g., first-course-start) to gain better insights into drop-off rates before L1 completion.

## 9.0 Final Dataset Summary

The dataset is now optimized and ready for deployment into the Business Intelligence environment.

| **Metric** | **Value** |
| --- | --- |
| **Rows (after cleaning)** | **66,704** |
| **Columns (after cleaning)** | **53** |
| **Unique users** | **69,188** |
| **Nulls remaining** | 267,788 (informative only) |

**Key columns ready for BI:** account\_created\_ts, user\_id, total\_points\_earned, num\_li\_completions, user\_rank\_number, verification\_status.