Initial Observations:

1. User Table (`USERS`)

- **Data Quality Issues:**
  - Missing values:
    - `BIRTH_DATE` (4% missing)
    - `STATE` (5% missing)
    - `LANGUAGE` (31% missing)
    - `GENDER` (6% missing)
  - `ID` (user identifier) appears to be a string, which is expected.
- **Potential Challenges:**
  - `BIRTH_DATE` and `CREATED_DATE` are stored as strings instead of date format.
  - The `LANGUAGE` field has many missing values, which could impact segmentation.
  - The `GENDER` field has 11 unique values, including non-standard entries

2. Products Table (`PRODUCTS`)

- **Data Quality Issues:**
  - Missing values:
    - `CATEGORY_1` (~0.01% missing)
    - `CATEGORY_2` (~0.17% missing)
    - `CATEGORY_3` (7% missing)
    - `CATEGORY_4` (92% missing – might not be a crucial field)
    - `MANUFACTURER` and `BRAND` (27% missing)
    - `BARCODE` (~0.48% missing)
  - `BARCODE` is stored as a floating-point number, which might cause precision issues.
- **Duplicate Records**
  - 215 duplicate records found in the `PRODUCTS` table.
- **Potential Challenges:**
  - The hierarchical categories (`CATEGORY_1`, `CATEGORY_2`, etc.) are not always filled, which could impact product classification.
  - Missing barcodes may prevent linking some products to transactions.
  - Standardize product names and categories (“Coca-Cola 12 Pack” vs “Coke 12pk”)

3. Transactions Table (`TRANSACTIONS`)

- **Data Quality Issues:**
  - `BARCODE` is missing for ~12% of records.

    - `FINAL_QUANTITY` has values like `"zero"` instead of numerical values, which will need cleaning.
    - `FINAL_SALE` sometimes appears empty.
    - Date columns (`PURCHASE_DATE` and `SCAN_DATE`) are stored as strings.
- **Duplicate Records**
    - 171 duplicate records found in the `TRANSACTIONS` table.
- **Potential Challenges:**
    - Transactions with missing `BARCODE` cannot be linked to products.
    - The `"zero"` value in `FINAL_QUANTITY` needs to be converted to numeric (possibly 0).
    - `PURCHASE_DATE` needs to always be on or before `SCAN_DATE` and were stored as strings
    - `RECEIPT_ID` has high cardinality (24,440 unique values for 50,000 entries), suggesting duplicates or multi-item receipts

Key Relationship Issues:

- **17,603 (35%)** user_id's in transactions **are not mapped to id's** in the user table which suggests incomplete or inconsistent user data
- **4,465 (41%)** barcodes in `TRANSACTIONS` **are not mapped to barcodes** in the products table which could stem from missing products in the catalog, discrepancies in barcode formats or data entry issues

Next Steps

- Remove duplicates from all tables.
- Timezone differences were handled by standardizing to datetime format.
- Convert `FINAL_QUANTITY` and `FINAL_SALE` to the correct numeric types
    - 12,500 non-numeric values
- Remove transactions with unmapped USER_IDs and BARCODEs.
- Standardize text fields (STORE_NAME, BRAND, MANUFACTURER).
- Make sure BARCODE is a valid numeric field.
- Convert date columns into datetime format
- Remove transactions where PURCHASE_DATE is after SCAN_DATE
- Make sure BIRTH_DATE is realistic (not in the future or unrealistic ages)
- Output a summary of the cleaned data.