Hello ,
After conducting an exploratory analysis on the Receipts, Users, and Products datasets, I came across several data quality issues that I believe are important to address. Here’s a summary of the key findings:

1. **Date and Timestamp Information**:
   - In the User and Transaction datasets, fields such as CREATED_DATE, BIRTH_DATE, PURCHASE_DATE, and SCAN_DATE include timezone information in inconsistent formats, and there are approximately 4% of missing birthdates. This inconsistency may affect any date-based analyses, such as tracking user sign-ups or purchase timelines accurately.
2. **Demographic and User Profile Information**:
   - **Language** and **Gender** data for users have a substantial amount of missing values (30% and 6%, respectively). Additionally, the GENDER field contains many variations (e.g., "non_binary," "Non-Binary," "not_listed," etc.), which makes consistent gender categorization difficult without standardization. This limits our ability to accurately segment users based on demographic information.
3. **Transaction Data and Product Identification**:
   - In the Transaction dataset, **BARCODE** information is missing in about 12% of records, which limits our ability to match products to transactions reliably. Additionally, FINAL_QUANTITY includes inconsistent values like "zero" instead of numeric values, potentially affecting the accuracy of transaction and inventory analyses.
   - In the Products dataset, over 90% of entries are missing CATEGORY_4 values, and a significant number of entries have placeholder values in MANUFACTURER. Missing or placeholder data reduces the accuracy of product categorization and affects any analyses requiring detailed product classifications.
4. **Financial and Purchase Information**:
   - The **FINAL_SALE** field in the Transaction dataset includes missing values in some transactions, meaning we lack information on the total sales amount for these receipts. This, combined with missing BARCODE values, impacts our ability to calculate revenue metrics accurately and tie specific purchases to products.
5. **Brand Classification and Categorization**:
   - The Products dataset lacks complete categorization in many entries, with CATEGORY_1 to CATEGORY_3 containing gaps in classification. Additionally, around 25% of BRAND and MANUFACTURER data is missing or contains placeholder text. These gaps complicate brand-level analyses, particularly when trying to identify popular brands or conduct market segment evaluations.

These issues affect our ability to accurately analyze user behavior, product performance, and brand engagement across different segments. To address these, I recommend prioritizing the standardization of gender values, consistent capture of barcodes, and enhanced completeness of product categorization fields. This will improve the accuracy and reliability of future analyses based on this data.

Let me know if you’d like to discuss these findings in more detail or if a meeting with the data engineering team would be helpful.

Best regards,
Dharnidhar Reddy Banala