Identified Data Quality Issues

1. **Null or Missing Values**
   a. Users Table:
      - Missing or null values in critical fields such as state, createdDate, and lastLogin.
      - Example: Some users might not have a state value, which can affect regional analysis.
   b. Receipts Table:
      - Fields like bonusPointsEarnedReason are often null.
      - Missing values in purchaseDate or totalSpent could lead to incomplete transaction records.
   c. Brands Table:
      - Missing or null values in the brandCode and categoryCode fields.
      - Example: Some brands have an empty brandCode.
2. **Inconsistent Data Types**
   a. Receipts Table:
   - Dates like createDate, purchaseDate, and pointsAwardedDate might be stored as strings instead of proper date/datetime types.
   - Example: "purchaseDate": "2025-02-15T12:34:56Z" should be standardized as a datetime object.
   b. Brands Table:
   - The barcode field might have inconsistent formats (e.g., numeric vs. string).
3. **Duplicate Records**
   a. Receipts Table:
   - Duplicate receipt entries with the same _id.
   - Example: Multiple records with the same receipt ID could inflate transaction counts.

   b. Brands Table:
   - Duplicate brand names with different _id values.
   - Example: Variants of "test brand" appear repeatedly with slightly different names and barcodes.
4. **Outliers or Invalid Values**
   a. Receipts Table:
   - Negative or zero values in fields like totalSpent or purchasedItemCount.
   - Example: A receipt showing "totalSpent": -50.00 is invalid.
   b. Brands Table:
   - Unrealistic barcodes, such as those containing non-numeric characters.
   - Example: "barcode": "511111abc123" is invalid for a barcode field.
5. **Referential Integrity Issues**
   a. Receipts Table:

- Some receipts reference non-existent users (userId) in the Users table.
- Example: A receipt with "userId": "nonexistent_user" cannot be joined properly.

b. Brands Table:

- Some brands reference invalid categories (categoryCode) that do not exist in the schema.
- Example: "categoryCode": "INVALID_CATEGORY" does not match any defined category.

Recommendations to address Data Quality Issues by specifying Metadata effectively, conforming data to Business rules and Data Visualization

1. **Null Value Handling**:
    - Enforce constraints during data ingestion to reject records with missing critical fields.
    - Use default values where applicable (e.g., set unknown states to "Unknown").
2. **Standardize Data Types**:
    - Convert all date fields to proper datetime types during ETL processes.
    - Ensure numeric fields (e.g., barcodes) are validated for consistent formats.
3. **Deduplication**:
    - Implement deduplication logic during data ingestion using primary keys or unique constraints.
4. **Outlier Detection**:
    - Apply validation rules to reject negative or unrealistic values at the point of entry.
5. **Referential Integrity Enforcement**:
    - Use foreign key constraints to ensure valid references between tables (e.g., Users and Receipts).
6. **Data Audits**:
    - Schedule regular data quality audits to identify and resolve issues proactively.

Note: Refer to the other document for SQL queries used to identify the data issues above.