Data Quality Issues Report

Date: 3/2/2025
Prepared by: Ramya Ravikanti

1. Overview of Data Quality Issues

This report highlights the data quality issues observed across several tables in the database, including users, receipts, brands, and receiptitems. The issues found range from missing values to potential data integrity concerns, which could impact the accuracy of business insights and decision-making.

2. Missing Values

Users Table:

- **Missing Fields:**
  - **LastLogin:** 40 records missing.
  - **SignUpSource:** 5 records missing.
  - **State:** 6 records missing.
- **Impact:** Missing LastLogin could affect our ability to track user engagement and retention. Missing SignUpSource and State could hinder analysis of user acquisition channels and geographies.

Receipts Table:

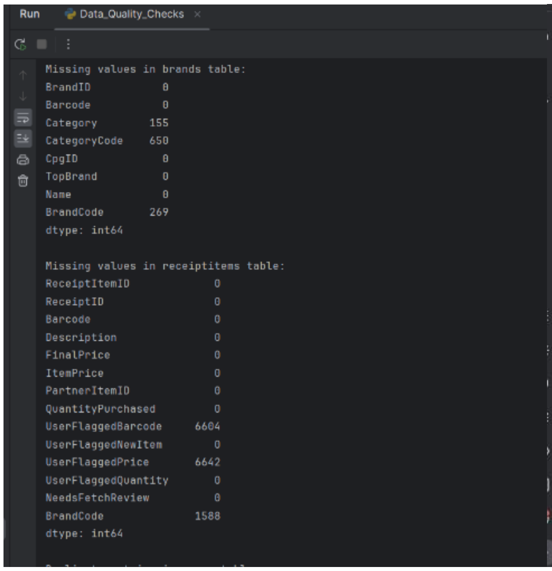
- **Missing Fields:**
  - **BonusPointsEarnedReason:** 575 records missing.
  - **FinishedDate:** 551 records missing.
  - **PointsAwardedDate:** 582 records missing.
  - **PurchaseDate:** 448 records missing.
- **Impact:** These missing values might impact the completeness of financial and rewards-related reports, especially for bonus points or transaction completion tracking.

Brands Table:

- **Missing Fields:**
  - **Category:** 155 records missing.
  - **CategoryCode:** 650 records missing.
  - **BrandCode:** 269 records missing.
- **Impact:** Missing values in **Category**, **CategoryCode**, and **BrandCode** can affect our ability to categorize products and conduct inventory or marketing analysis effectively.

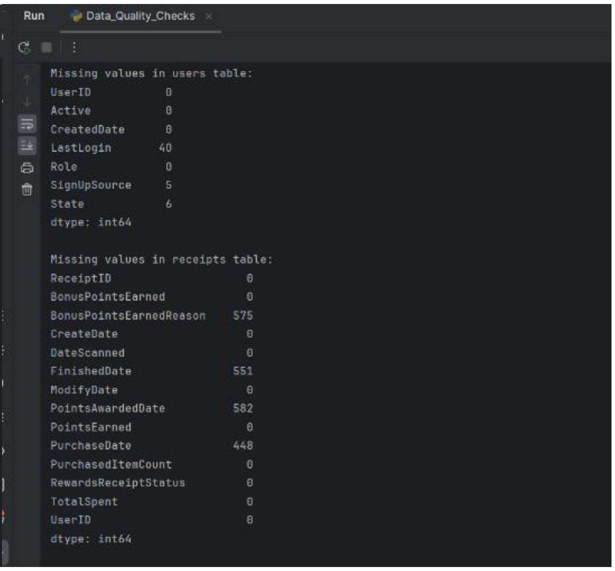
ReceiptItems Table:

- **Missing Fields:**
  - **UserFlaggedBarcode:** 6604 records missing.
  - **UserFlaggedPrice:** 6642 records missing.
  - **BrandCode:** 1588 records missing.
- **Impact:** Missing flagged data and **BrandCode** could effect quality checks on receipt items, affecting the integrity of item validation and reporting.



```
Missing values in brands table:
BrandID           0
Barcode           0
Category        155
CategoryCode    650
CpgID             0
TopBrand          0
Name              0
BrandCode       269
dtype: int64

Missing values in receiptitems table:
ReceiptItemID            0
ReceiptID                0
Barcode                  0
Description              0
FinalPrice               0
ItemPrice                0
PartnerItemID            0
QuantityPurchased        0
UserFlaggedBarcode    6604
UserFlaggedNewItem       0
UserFlaggedPrice      6642
UserFlaggedQuantity      0
NeedsFetchReview         0
BrandCode             1588
dtype: int64
```



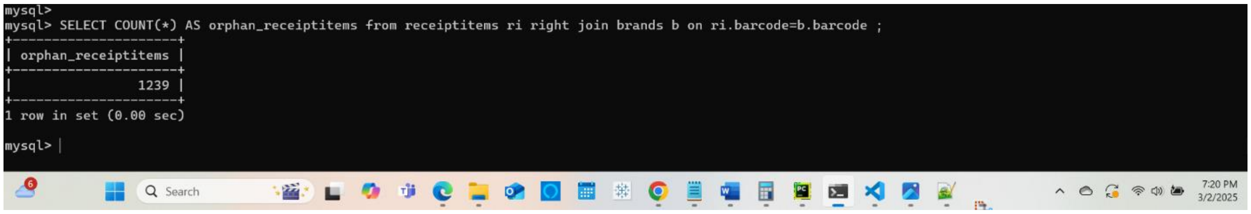
```
Missing values in users table:
UserID           0
Active           0
CreatedDate      0
LastLogin       40
Role             0
SignUpSource     5
State            6
dtype: int64

Missing values in receipts table:
ReceiptID                  0
BonusPointsEarned          0
BonusPointsEarnedReason  575
CreateDate                 0
DateScanned                0
FinishedDate             551
ModifyDate                 0
PointsAwardedDate        582
PointsEarned               0
PurchaseDate             448
PurchasedItemCount         0
RewardsReceiptStatus       0
TotalSpent                 0
UserID                     0
dtype: int64
```

3. Orphaned Records

Orphaned ReceiptItems records:

- **Issue:** 1,239 ReceiptItems records have a barcode that doesn't match any entry in the brands table.
- **Impact:** These orphaned records likely represent untracked products or errors in the barcode assignment process, which could lead to incorrect or incomplete inventory data.
- The query used to extract data on brand performance for the most recent and previous months may yield empty results if a direct INNER JOIN is used between receiptitems and brands, because of the missing barcode information.

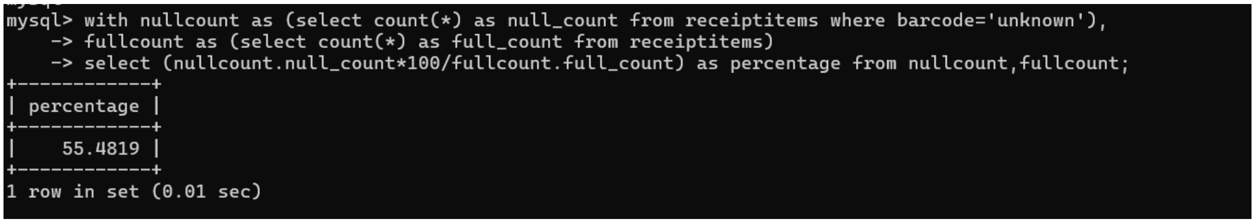


```
mysql>
mysql> SELECT COUNT(*) AS orphan_receiptitems from receiptitems ri right join brands b on ri.barcode=b.barcode ;
+---------------------+
| orphan_receiptitems |
+---------------------+
|                1239 |
+---------------------+
1 row in set (0.00 sec)

mysql> |
```

4. Missing Barcodes in ReceiptItems Table

- **Issue:** 55.48% of records in the receiptitems table have a missing barcodes.
- **Impact:** This could lead to inaccurate tracking of items sold, which can affect both sales and inventory reporting.



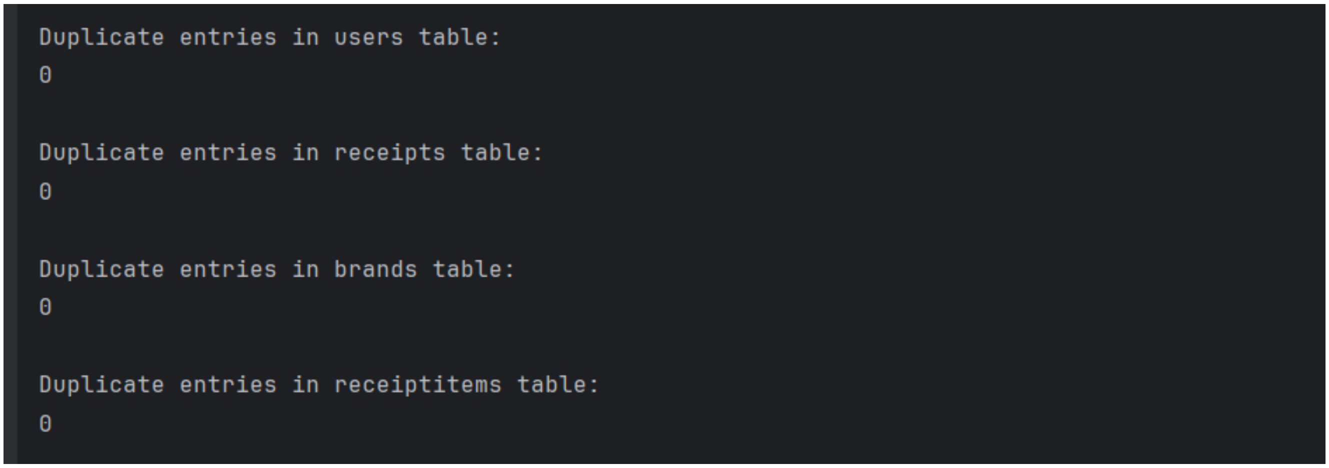
```
mysql> with nullcount as (select count(*) as null_count from receiptitems where barcode='unknown'),
    -> fullcount as (select count(*) as full_count from receiptitems)
    -> select (nullcount.null_count*100/fullcount.full_count) as percentage from nullcount,fullcount;
+------------+
| percentage |
+------------+
|    55.4819 |
+------------+
1 row in set (0.01 sec)
```

5. Outliers

- **Issue:** No outliers were identified in the TotalSpent and QuantityPurchased columns of the receipts table based on negative values.
- **Impact:** No outliers were found.

6. Duplicate Entries

- **Users Table:** No duplicate entries found.
- **Receipts Table:** No duplicate entries found.
- **Brands Table:** No duplicate entries found.
- **ReceiptItems Table:** No duplicate entries found.
- **Impact:** Clean records with no duplicates ensure data integrity, making it easier to perform accurate analysis.



```
Duplicate entries in users table:
0

Duplicate entries in receipts table:
0

Duplicate entries in brands table:
0

Duplicate entries in receiptitems table:
0
```