Overview

This documentation briefly discusses data quality assessment on the three datasets provided - Receipts, Users and Brands. The goal of the assessment is to provide a starting point for an exhaustive data quality check. That being said, while still elementary, these checks are a good starting point for understanding the data quality.

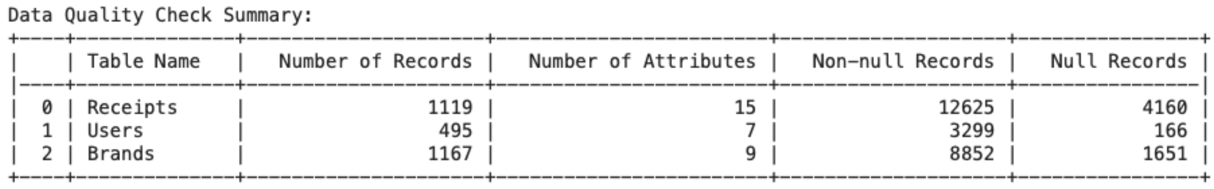
This will help us to highlight principal areas for improvement and sets the stage for more comprehensive data quality management efforts for the future.

Data Quality Dimensions

The assessment focuses on 6 key dimensions-

1. Data Types - Understanding the original data types in the data set and for the sake of the assessment issuing appropriate conversions.
2. Completeness - Completeness checks help us identify the presence of missing values. Attributes that have significant missing values are - bonusPointsEarned, bonusPointsEarnedReason, and pointsAwardedDate in the Receipts dataset, and state in the Users dataset.
3. Validity - an important aspect of data quality is its validity. Attribute values are supposed to conform to their domain ranges. For example, if we are referring to dateScanned attribute, its values are supposed to be of date time format, anything else would then be a non-domain value.
4. Consistency - these checks are to ensure that the data values are logically consistent in the dataset. For the preliminary test done, there were no issues reported.
5. Integrity - integrity checks ensure the referential integrity is valid between the relationships. No integrity issues were found
6. Outlier Detection - I used both IQR and Z-score methods to identify any outliers in the numeric attributes. Attributes such as pointsEarned, purchasedItemCount, and totalSpent in the Receipts dataset have outliers and may require further analysis.

Summary of the checks:



```
Data Quality Check Summary:
+----+--------------+---------------------+------------------------+--------------------+----------------+
|    | Table Name   |   Number of Records |   Number of Attributes |   Non-null Records |   Null Records |
|----+--------------+---------------------+------------------------+--------------------+----------------|
|  0 | Receipts     |                1119 |                     15 |              12625 |           4160 |
|  1 | Users        |                 495 |                      7 |               3299 |            166 |
|  2 | Brands       |                1167 |                      9 |               8852 |           1651 |
+----+--------------+---------------------+------------------------+--------------------+----------------+
```

Key Findings

Receipts Dataset

Attribute	Issue	Impact
bonusPointsEarned	51.39% missing values	Significant impact on reward analysis; data imputation or alternative strategies required
bonusPointsEarnedReason	51.39% missing values	Affects understanding of reward mechanisms; needs review for accurate reward insights
pointsEarned	Stored as object instead of numeric	Impedes numerical analysis; conversion to appropriate data type necessary
purchasedItemCount	43.25% missing values	Affects transaction completeness; critical for accurate purchase analyses
rewardsReceiptItemList	No missing values	No issues found
rewardsReceiptStatus	No missing values	No issues found
totalSpent	Stored as object instead of numeric; 38.87% missing values	Affects financial analyses; conversion to numeric and imputation required
finishedDate.$date	49.24% missing values	Impacts time-based analyses; needs imputation or alternative date handling
pointsAwardedDate.$date	52.01% missing values	Affects reward point tracking; requires imputation or review
purchaseDate.$date	40.04% missing values	Important for purchase trend analysis; needs imputation

Users Dataset

Attribute	Issue	Impact
active	No missing values	No issues found
role	No missing values	No issues found
signUpSource	9.70% missing values	Affects user acquisition analysis; requires imputation or source clarification
state	11.31% missing values; invalid state codes (multiple instances of 'nan')	Impacts geographic analysis; needs standardization and imputation
createdDate.$date	No missing values	No issues found
lastLogin.$date	12.53% missing values	Affects user engagement tracking; needs imputation

Brands Dataset

Attribute	Issue	Impact
barcode	No missing values	No issues found
category	13.28% missing values	Affects product categorization; needs imputation or category assignment
categoryCode	55.70% missing values	Significantly affects categorization accuracy; requires substantial imputation or review
name	No missing values	No issues found
topBrand	52.44% missing values	Affects identification of top brands; requires review and potential imputation
brandCode	20.05% missing values	Affects brand identification and integrity; needs imputation or review

Outlier Detection Summary

Table Name	Attribute	IQR Outlier Count	Z-score Outlier Count
Receipts	bonusPointsEarned	0	0
Receipts	pointsEarned	36	17
Receipts	purchasedItemCount	43	15
Receipts	totalSpent	55	7

Recommendations

1. **Handling Missing Values** - Implementing imputation strategies for attributes with missing values. For wherever it is not possible, coming up with an alternative solution such as setting default values.
2. **Validating and Correcting data** - Deploying validation scripts and checks to ensure invalid entries do not make way to the warehouse. Constraints and other checks can also be added to the database schemas.
3. **Improving Data Consistency** - Employing data consistency rules to ensure there is no logical fallacy in the datasets, especially for attributes that are shared across tables.
4. **Outlier Management** - Consistently check for outliers in the data, ensure pipelines have scripts in place to ensure garbage values do not pass through, ultimately ensure what exactly these outliers indicate - data entry errors or legitimate extreme values.
5. **Data Audits and Monitoring** - develop monitors to continuously monitor, maintain and report any data quality issues.
6. **Robust ETL processes -** Pipelines must be designed with validation checks are integrated to ensure data accuracy and efficiency. Define transformation rules for proper data structure conversion.

Some other recommendations -

1. Invest in advanced data quality solutions
2. Implement MDM practices
3. Provide comprehensive training for staff involved with data pipelines, processing and entry.
4. Establish a data governance framework company wide to ensure there are no gaps in understanding of metadata.

Conclusion

While this is not a comprehensive, 360 check for data quality issues, this, as mentioned previously, should serve as a good starting point along with the recommendations to issues found.