Gmail

Gandhi Edhala <gandhijsg04@gmail.com>

Data Quality Issue and Next Steps

Gandhi Edhala <gandhijsg04@gmail.com>
Draft To: gandhinaidu04@gmail.com

Sat, Mar 8, 2025 at 7:34 PM

Hi,

I've reviewed the provided datasets (`users` , `receipts` , and `brands`) and identified several data quality issues that need attention. Below is a summary of the issues and a few questions to help resolve them.

Data Quality Issues

1. **Missing Values**:
   - `Users` : Missing `lastLogin` , `signUpSource` , and `state` .
   - `Receipts` : Missing `bonusPointsEarned` , `finishedDate` , and `totalSpent` .
   - `Brands` : Missing `category` , `categoryCode` , and `brandCode` .
2. **Duplicates**: 283 duplicate user records.
3. **Invalid Relationships**:
   - 148 invalid `userId` values in `Receipts` .
   - 7,299 invalid `barcode` values in `ReceiptItems` .
4. **Outliers**: Extreme values in `totalSpent` (e.g., a maximum of 1000).

Questions for You

1. **Missing Data**:
   - Should we remove records with missing values or fill them (e.g using defaults or averages)?
   - Are there additional sources to clarify missing values ?
2. **Invalid Relationships**:
   - Should we remove records with invalid `userId` or `barcode` values, or update them?
3. **Optimization**:
   - What additional data would help improve the analysis (e.g., user demographics, product details)?

4. **Scaling**:
    - Should we plan for performance improvements (e.g., indexing, partitioning, or moving to a distributed database)?

Next Steps

1. Clean the data by addressing missing values, duplicates, and invalid relationships.
2. Refine the analysis based on your feedback.

Let me know your thoughts or if there's anything else you'd like to prioritize. I'm happy to discuss further!

Best regards,
Gandhi Edhala