Subject: Data Quality Insights and Action Plan

Hi [Product/Business Manager's Name],

I wanted to share some initial insights from my analysis of the datasets we've been working with. Addressing these data quality issues is crucial for ensuring accurate and reliable analysis and decision-making going forward.

Data Quality Findings:

Null Values:

- Many columns across all three datasets contain null values.
- Particularly, the "brands" table exhibits over 50% null values in some columns, complicating preprocessing and analysis.

Duplicate Rows:

- The "users" table contains over 250 duplicate rows, potentially skewing our analysis results.

Receipt Table Issues:

- Critical columns related to brands are absent in the "receipt" table, making it challenging to perform brand-level analysis.
- Additionally, duplicate items within single receipts were observed, necessitating further investigation.

Questions and Next Steps:

Questions:

- Clarification on the business context of missing data in the "brands" table would aid in informed decision-making.
- Determining the best approach to link brands and receipts tables effectively is essential for comprehensive analyses.
- More context around the description of "rewardsReceiptItemList" would enhance our understanding of nested columns.

Discovery Process:

- These issues were identified through systematic review, focusing on null values, duplicates, and manual checks of all tables.

Resolution Needs:

- Strategies such as imputation for null values and removal of duplicate rows in the "users" table need to be decided upon.

- Consideration for creating a brand-related column in the "receipt" table is crucial for better integration and analysis.

Optimization Information:

- Understanding primary business objectives will guide optimization efforts for data assets.

Performance Concerns:

- Anticipated issues with dataset size, especially given a large number of null values, require efficient data-cleaning processes and query optimization.

Action Plan:

- Establish guidelines for handling missing values and implement deduplication processes.
- Investigate causes of data skewness and inconsistencies, especially in the "brands" table.
- Develop strategies for imputing missing values in the "receipts" table while maintaining data integrity.
- Define rules for categorizing and handling outliers to improve data reliability.
- Conduct regular audits to identify and rectify quality issues promptly.

By addressing these data quality issues, we can significantly enhance the accuracy and reliability of our analyses, leading to better-informed decision-making.

Your insights and guidance on the business context will be invaluable as we continue to optimize these data assets.

Please let me know if you have any questions or if there's any additional information you'd like to discuss. Looking forward to your feedback.

Best regards,
Rahul Maddula