Subject: Data Quality Issues Identified and Next Steps

Hello Stakeholder,

I hope this message finds you well. I am writing to inform you about some critical data quality issues we've identified during our recent analysis and to outline the next steps required to address these issues effectively.

Key Findings:

1. **High Null Values in Receipt Items:**
    - We discovered that approximately 27 out of 34 columns in the **receipt_items** dataset have more than 50% null values. This high level of missing data can significantly impact the accuracy of our analytics and the insights derived from this data.
2. **Inconsistent Brand Codes:**
    - There are 670 more unique brand codes in the **brands** dataset compared to the **receipt_items** dataset. Additionally, we identified 186 brand codes in the **receipt_items** dataset that are not present in the **brands** dataset.
3. **User Data Mismatch:**
    - We found that 148 user IDs present in the **receipts** dataset do not match any user IDs in the **users** dataset. This discrepancy suggests potential issues with data integration or user tracking.

Questions and Information Needs:

1. **Clarification on Data Collection:**
    - How are the receipt and brand data collected and integrated into our system? Understanding the source and process can help identify where data loss or corruption may occur.
2. **Consistency in Brand Codes:**
    - Is there a standardized method or codebook for assigning brand codes across different datasets? Ensuring consistency here is crucial for reliable data merging.
3. **User Data Validation:**
    - Can we obtain more information on how user data is tracked and merged with transaction data? This might help us understand the cause of the mismatches.

Next Steps to Resolve Data Quality Issues:

1. **Data Cleaning and Validation:**
    - We need to implement robust data cleaning processes to handle null values, such as imputation techniques or flagging incomplete records for further review.
    - Establish a validation mechanism to ensure all brand codes and user IDs are correctly mapped and consistent across datasets.
2. **Enhanced Data Integration:**
    - Review and potentially redesign our data integration pipeline to minimize data loss and improve accuracy. This might involve working closely with the IT and data engineering teams.
3. **Performance and Scaling Considerations:**
    - As we address these data quality issues, we must also consider the scalability of our solutions. The data cleaning and validation processes should be optimized for performance to handle large datasets efficiently.
    - Regular monitoring and automated alerts for data quality metrics can help maintain data integrity as we scale.

Additional Information Required:

- Detailed documentation on the current data collection and integration processes.
- Access to historical data for comparison and deeper analysis.
- Insights into any upcoming changes to data sources or structures that may impact our current work.

Addressing these data quality issues is crucial for ensuring the reliability and accuracy of our analytics. Your support in providing the necessary information and resources will be invaluable.
Please let me know if you need any further details or if we can discuss this in more detail in our next meeting.

Best regards,
Prinjal Dave