Subject: Data Quality Findings & Actionable Insights

Dear Stakeholders,

I hope you’re well!

I’m writing to share some preliminary insights from our analysis of the transactions, products, and users datasets, along with a few data quality concerns that we need to address to ensure the accuracy of our reporting.

Key Actionable Insight:

Our transaction data, spanning from June 2024 to early September 2024, reveals a clear pattern: receipt scans on weekdays consistently far outnumber those on weekends. This suggests that our users are significantly more engaged during the workweek.

One plausible explanation for this trend is the impact of the summer season. During this period, many students and employees may be taking vacations or enjoying a more relaxed weekend schedule, which reduces their likelihood of engaging with our scanning process on weekends. In contrast, weekdays tend to follow a more structured routine, prompting users to incorporate scanning into their daily habits—perhaps during commutes or while running errands.

This insight presents a valuable opportunity for us. By developing targeted weekend promotions or engagement strategies, we can stimulate user activity during the slower weekend period. For example, we might consider:

- Offering exclusive weekend rewards or contests.
- Launching special discounts available only on weekends.
- Creating campaigns that encourage users to share their weekend shopping experiences.

By addressing this weekend engagement gap, we can balance overall user activity and drive additional growth, ensuring that our platform remains active and attractive throughout the entire week.

Data Quality Findings:

Transactions Dataset:

- **Missing BARCODE Values (12%):**
  Records missing BARCODE values were removed to ensure accurate joins with the product dataset.
- **Duplicate Receipt IDs:**
  We identified 25,389 cases where duplicate receipt IDs had one row with a zero for FINAL_QUANTITY or FINAL_SALE and another with a valid value. These were merged by retaining the nonzero values.
- **Additional Issues:**
  - 171 exact duplicate rows were removed to prevent data inflation.
  - 94 records with SCAN_DATE preceding PURCHASE_DATE were removed due to probable data entry errors.
  - One receipt with an unrealistic FINAL_SALE value of 276 was removed as an outlier.

Products Dataset:

- **CATEGORY_4 Field:**
  Over 90% of records in this field are missing, suggesting that this level of product categorization is incomplete or inconsistently recorded. It could be excluded from our analysis.
- **Missing Manufacturer/Brand Info:**
  Approximately 26% of records lack values in the MANUFACTURER and BRAND fields, potentially affecting our brand-level insights.
- **BARCODE Integrity Issues:**
  3,968 records missing BARCODE values were removed, and duplicate BARCODE entries with conflicting BRAND values were flagged for further validation.

Users Dataset:

- **GENDER Inconsistencies:**
  Variations in gender entries (e.g., “non_binary” vs. “Non-Binary”) require standardization for accurate demographic segmentation.
- **Date Anomalies:**
  Records where CREATED_DATE precedes BIRTH_DATE have been removed.
- **Missing LANGUAGE Data:**
  Approximately 30% of user records are missing LANGUAGE values.

Outstanding Questions & Next Steps:

- **Duplicate Receipt IDs:** Should we always retain the nonzero FINAL_QUANTITY/FINAL_SALE values, or could zeros sometimes represent legitimate adjustments (e.g., returns or discounts)?
- **Outlier Transactions:** Should wholesale or bulk purchase receipts be included in our analysis, or should they be treated as anomalies?
- **Duplicate BARCODEs in Products:** Could these discrepancies be due to rebranding or data entry errors? What is the best approach to validate the correct product-brand associations?
- **Potential Duplicate Users:** How should we handle cases where multiple accounts might belong to a single individual?

Your guidance on these questions would be invaluable as we refine our analysis. Please refer to the attached Jupyter Notebook for a more detailed exploration.

Thank you for your time and support. I look forward to your feedback.

Best regards,

Vivian