**Subject:** Data Quality Insights & Next Steps for Review

Hi,

I hope you’re doing well! I wanted to share an update on the initial analysis we've conducted on our users, products, and transaction data. Over the past few days, we’ve explored the datasets to uncover key trends and potential areas for improvement. Below, I've summarized the main findings and highlighted some key questions where I’d appreciate your guidance.

**Key Data Quality Issues**

**Missing User Information:**  
We’ve observed that a significant portion of user profiles are incomplete, with key fields like birthdates, states, and genders often missing. Some examples include:

* **Birthdates** are frequently either missing or set to placeholder values across many records.
* **State information** is absent for a notable percentage of users, which could hinder regional analysis or targeted campaigns.
* **Gender data** is inconsistent or missing entirely, with redundant categories like "not listed" and "Prefer not to say" causing confusion.

These gaps are making it difficult to properly segment users or extract meaningful insights from the data.

**Product Data Inconsistencies:**  
There are several issues in the Products table, particularly around missing or placeholder values:

* **Categories** (especially CATEGORY\_3 and CATEGORY\_4) have a high percentage of missing data, with CATEGORY\_4 being empty in 92% of records.
* **Manufacturer and brand information** is often incomplete or replaced with placeholder values like "PLACEHOLDER MANUFACTURER."
* **Barcodes** are either missing or duplicated, potentially causing mismatches between products and transactions.

These issues limit our ability to analyze product performance and manage inventory effectively.

**Transaction Data Issues:**  
The Transactions table has some key challenges:

* **Duplicate records** are appearing, particularly with RECEIPT\_IDs where some entries vary only in FINAL\_QUANTITY or FINAL\_SALE. This could point to data entry errors or system glitches.
* There is a **barcode mismatch** between the Transactions and Products tables, with only 59% of barcodes matching. This inconsistency could affect product-level reporting and inventory tracking.
* The **transaction date range** is limited to just 89 days in 2024, which might not be enough for long-term trend analysis.

**User ID Discrepancy:**  
We also identified a discrepancy between the number of users in the Users table and those in the Transactions table. Specifically, only 91 users from the Transactions table are represented in the Users table, which could indicate either incomplete data in the Users table or a disconnect in how user data is being captured and linked.

**Notable Trend**

Despite data quality challenges, our analysis highlights a clear trend: Tostitos dominates the dips and salsa category, leading in both revenue and transaction volume. With a 57% greater share of receipts than its closest competitor, Pace, Tostitos has solidified its position as the category leader, reflecting strong consumer preference and brand loyalty.

**Next Steps & Questions**

To address these issues and ensure that our data is both accurate and actionable, I’d like your input on the following:

**Users Table:**

* What’s causing the frequent missing data in fields like birthdates, states, and genders? Can we make these fields mandatory during user registration or offer incentives for users to provide this information?
* How can we standardize gender categories to eliminate redundancies (e.g., combining "not listed" and "Prefer not to say")?
* What’s the root cause of the discrepancy between the Users and Transactions tables? Is this a data integration issue, a system error, or something else?

**Products Table:**

* Why is category, manufacturer, and brand data often missing? Can we enrich this data externally or improve our internal data collection processes?
* What’s the reasoning behind placeholder values like "PLACEHOLDER MANUFACTURER"? Should we replace these with accurate information or standardize them to "Unknown"?
* How are barcodes assigned and validated? Can we implement stricter validation rules to prevent missing, duplicate, or mismatched barcodes?

**Transactions Table:**

* What’s the process for handling duplicate RECEIPT\_IDs, especially when they differ in FINAL\_QUANTITY, FINAL\_SALE, or barcodes? Should we remove these duplicates or flag them for review?
* Why are only 91 users from the Transactions table represented in the Users table? Are there users in the Transactions table who are not captured in the Users table, and if so, why?
* How can we improve the match rate between transaction barcodes and product barcodes? Is there a process in place to ensure consistency across these tables?

**Moving Forward**

I’d love to schedule some time to dive deeper into these findings and discuss the next steps. Let me know what works best for your schedule, and I’ll set up a meeting. In the meantime, feel free to share any thoughts or questions you have.

Looking forward to hearing from you!

Best,

Nidhi