Subject: Summary of Data Quality Review & Insights – Request for Follow-Up

Hi Todd,

I wanted to provide a brief overview of the user, transaction, and product datasets that I recently analyzed. I found a number of data quality problems during the review process that, if left unchecked, might affect analysis and business choices.

Key Issues:

- **Missing values:** A lot of users lack language, state, and birthdate information. A number of product-related records lack category, manufacturer and brand information. Additionally, some barcodes and sale amounts are missing from transactions.
- **Inconsistent formatting**: String values like "zero" are used in place of numeric zero in certain numeric fields, such as quantity and sale. Additionally, barcode values have a mix of numeric and scientific notation which has an impact on how tables relate to one another.
- **Duplicates:** There are numerous duplicate barcodes and completely duplicated rows in the product file. Duplicate receipt IDs are also present in the transaction file; some of them might be valid multi-line items, while others would require more investigation.

Interesting Trend: Store Drives Scan Delay More Than Age

I compared average scan delays by both age group and store:

- While users aged 30–60 show slightly higher delays (averaging just over 3 days), the delay gap across age groups is relatively small.
- In contrast, certain stores—notably Vitacost—have average scan delays of nearly 20 days, far exceeding any age-based pattern.

This suggests that store-specific factors—such as receipt formatting, reward eligibility, or purchase context—are likely more influential in delaying receipt scans than demographic characteristics alone.

Request for Action:

To help resolve these data questions and build toward more reliable insights, I could use a bit more input:

1. **Barcode consistency:** Should all barcodes be treated as strings in the long term? If so, we'll need to standardize that across files.
2. **Receipt scanning behavior:** Should we treat long scan delays at certain stores as expected behavior, or should they trigger a review or follow-up?
3. **Product metadata:** Is there a more complete product catalog (with brand/category hierarchy) that we can link to for missing or duplicate product records?

Let me know if you'd like to connect for a quick sync, or I'm happy to refine these findings further based on your priorities.

Thank you.