Hello,

I hope this email finds you well. I'm currently examining the datasets we are working with, and I've identified some data quality issues that require clarification to resolve effectively. Additionally, I'd like to better understand the broader objectives for these data to ensure alignment with business goals.

1. Understanding the Data

a. Where does the data come from, and what transformations has it undergone?
b. What is the primary purpose of this dataset (e.g., analytics, reporting, decision-making)?
c. Are there specific fields or metrics that are important for our goals?
d. Could you clarify the meaning of fields in rewardsReceiptItemList?

2. Data Quality Issues

During my exploratory analysis of the three datasets, I discovered the following major issues in the **Brands** data:

a. Some brands lack category information, which may affect analysis by product segments.
b. There are overlaps in category names, such as "Beauty," "Personal Care," and "Beauty & Personal Care.", which could lead to duplicate or inaccurate reporting. We may consider combining or removing some categories.
c. Some brand names are placeholder names like "test brand…", which may need filtering or cleanup.
d. There are sub-brands listed under the same brand (e.g., ONE A DAY®, ONE A DAY® 50+, ONE A DAY® KIDS). Should these sub-brands be consolidated under their parent brand, or treated separately?

In **Users** data:

a. Missing values in the users' last login date: Does this indicate they never logged in after registration?

In **Receipts** data:

a. Missing values in many fields, such as bonus points earned and total spent.
b. The targetPrice field in rewardsReceiptItemList has only two unique values, 77 and 800, which seems counterintuitive. Could you please provide clarification?

Resolving these data quality issues is important for getting accurate and actionable insights. I'd appreciate any guidance or solutions you can provide on these points.

3. Performance and Scaling Considerations

As the datasets grow, I foresee the following challenges:

a. Storage and Retrieval: Larger data volumes may lead to storage capacity issues or slower response times, impacting reporting efficiency. **Proposed Solution**: Use flexible, scalable cloud-based storage solutions that can adapt as the data grows while keeping costs optimized.
b. Data Processing: Current workflows may take significantly longer to handle larger datasets. **Proposed Solution**: Transition to a system that processes data incrementally as it becomes available, instead of waiting for large batches. This will help maintain consistent processing speeds.

Your insights on these issues will help ensure we address data quality effectively and prepare for future growth. Please let me know if you'd like to discuss this further.

Thank you for your time and guidance!