Hi (name),

Upon analyzing anomaly patterns in our **users**, **brands**, and **receips** data using Python, I identified several data quality issues that could impact analytics accuracy and resource allocation efficiency:

1. **Duplicates in Users Data:** Approximately 60% of user records are duplicates, which inflates metrics and misrepresents user behavior.
2. **Test Data in Brands:** There are test records in the brands dataset (~37%) that need to be excluded to maintain accuracy.
3. **Many IDs Associated with a given Brand Name:** There are 11 brand names that don’t have a unique brand ID which could lead to inconsistent or wrong insights been drawn.
4. **Sparse Data in Receipt Items:**
   * Over 50% of the columns have more than 90% missing values
   * Our data connections to our **brands** table is also significantly underpopulated (only ~30% of these key connecting elements are available).
5. **Category Discrepancy:** The **category** field is better populated (13% missing) than **categoryCode** (56% missing), while **categoryCode** is presumably meant to be just the upper case of **category** with special characters removed.

**Questions and Next Steps:**

* What are the critical use cases for this data, and what are some of the must-haves for our operations and analytics?
* Are there known gaps in data collection or upstream processing that could explain these issues?
* Is there a team that’s in charge of the data validation piece that I should reach out to and work together with to address the issues and enhance the process?
* Are there additional data sources or documentation for our design and processes so that I can better understand the intent behind underpopulated fields like **brandCode** and **barcode**?

**Scaling and Performance Considerations:**  
If these data quality issues persist, they may introduce and amplify inaccuracies in our reporting, reduce analytics efficiency and generate redundant costs. (For example, we would not be able to accurately show business metrics for a significant number of brands, and our $1000 bill for user data storage could have been a $400 bill).

I propose implementing tests and validation processes in upstream systems if possible; and implementing a “staging” layer in our data warehouse for data cleansing if some of these tests are challenging to set up in upstream systems. We also need to implement deduplication logic, and clearer field requirements to ensure reliability and scalability of our systems.

Do you have 20-30mins later this week or early next week to hop on a quick follow-up call to align on priorities, data usage, and next steps? Please let me know a time that works the best for you!

Thanks,

Mark