Subject: Summary of Data Issues and Next Steps

Hi Team,

Hope you are doing well!

I analyzed our core user, brand, and receipt data sets and wanted to highlight the key data quality issues:

- Duplicate user records exist, making the data inconsistent for analysis. Deduplication is needed.
- Missing and incorrect values are present for user attributes like signUpSource and state.
- Brand codes referenced in receipts are not defined in the brands master list. This leads to incomplete brand affiliation.
- Critical brand attributes like category and brandCode are missing for hundreds of records.

I discovered these gaps by profiling the data - analyzing statistically for duplicates, validating values, and joining tables to inspect inconsistencies.

To resolve this, we need to:

- Deduplicate users table to create one master source of truth
- Perform lookup between receipts and brands to populate missing brand codes
- Fix incorrect nan/null values in user metadata
- Add missing category, categoryCode, brandCode for all brands

For scalability, normalized DB schemas with indexed fields for frequent filters and joins is crucial. We can also cache aggregated data.

Please let me know if you need any other details on the data issues or potential solutions. I'm ready to partner with you and the product team to fix these gaps and optimize our data assets.

Best regards,
Kedar Takwane