

Subject: Concerns regarding the Data Quality Issues

Hello [Business Stakeholder name],

Hope you are doing well. I am writing this email to ask you a few data quality issues that I found while I was exploring the datasets and building out the new data model. Some of them have been fixed while the rest needs your domain expertise. They are as follows:

- I analyzed users, brands and receipts files in which I found multiple columns with no values. Few of the columns like "categoryCode" and "brandCode" in the brands file have been filled using "category" and "name" respectively. Based on my analysis, these fields were similar and held no risk in data inaccuracy. Do you have any concern with that approach? If so, I can revert that and work on other alternative.
- For the other columns, I have added the tag "unknown" or kept it null. We will have to pair on those since I couldn't use any other columns to fill them up and I didn't want to add any misleading info.
- I found about 148 users that have a record in the receipts data but are not present in users file. There could be an issue in the way we are loading user info into the user file. I can reach out to back-end engineering team to investigate further but I thought of confirming with you in case not having these user's info in the users file was intentional.
- Also, there are multiple records for certain user ids in the users file. This could again be an issue with data loading process. Ideally, each user is supposed to have a single entry in the users dataset.

Fortunately, the volume we are dealing right now makes it easier to manage but as the platform scales up, these inconsistencies can cause major issue with the way we conduct analyses.

I am happy to schedule a 1 hour call for the 2 of us to talk more about this. Please let me know when you are available so we can connect.

Regards,

Rahul Nair