

Subject: Data Quality Update and Next Steps

Hello Team lead,

I hope you're doing well. I wanted to share an update on our recent work involving data quality analysis and optimization. I'll make sure to keep things clear and easy to follow.

Questions About the Data: As we continue to dive into the data, I'm particularly interested in understanding the "rewardsReceiptItemList" field from the "receipt.json.gz" file. Gaining more insight into its significance will help us align our analysis with your strategic goals more effectively.

Discovery of Data Quality Issues: During our exploration using Python notebooks, we've pinpointed several data quality concerns that are worth addressing:

- **Missing Values:** We've identified missing values in the "user," "receipt," and "brands" files that need attention.
- **Duplicate Records:** There are instances of duplicated records in the user file, which might skew our analysis results.
- **Role Inconsistency:** The "role" column should ideally have a consistent value of "consumer," but we found some instances with the value "fetch-staff."
- **UserID Mismatches:** We've noticed that some userIDs mentioned in the receipt file don't match any userIDs in the user file.

Resolution and Optimization: To tackle these data quality issues, we're planning to collaborate with the team to gain a better understanding of the data. Our next steps involve data cleaning and preparation before we integrate it into the database. This will ensure that the data we work with is accurate and reliable.

Optimization Insights: For optimizing the data assets we're creating, it would greatly help to understand the specific insights you're seeking to derive. Knowing your expectations will guide us in tailoring our efforts to generate the most valuable outcomes.

Thank you for your time and support. Looking forward to our continued collaboration.

Best regards,

Rohit Chavan

rdchavan.01993@gmail.com