**Subject: Addressing Critical Data Quality Issues for Accurate Decision-Making**

Hi Product/Business Leader,

I hope this message finds you well. As part of our ongoing efforts to optimize our data assets and ensure the integrity of our analyses, we have conducted a comprehensive review of our current datasets (relating to receipts,users and brand). This has led to the identification of several critical data quality issues that require immediate attention to support informed decision-making and strategic planning.

Upon analysis, we discovered that the brand with the highest spend among users created in the last six months is recorded as 'None', with Pepsi being the second highest. Similarly, 'None' is also recorded as the brand with the most transactions, again with Pepsi following. This inconsistency in data labeling prevents us from accurately attributing sales and transactions to specific brands, posing a significant barrier to our marketing and sales strategies.

Further examination revealed key issues in our users' data, including:

- Approximately 13% of lastLogin data is missing, which could significantly impact our marketing strategy's effectiveness.

- Around 11% of State information is missing, potentially affecting our ability to allocate marketing spend effectively by state.

- The user\_id field is not unique, which may lead to inaccurate querying and reporting.

In the brand data, we found:

- About 20% of brandCode data is missing, hindering our ability to correctly attribute sales and transactions.

- A staggering 52% of topBrand information is lacking, which is crucial for identifying top-performing brands.

- Additionally, categoryCode and category fields are missing approximately 55% and 13% of their data, respectively, affecting our marketing segmentation capabilities.

To address these issues and improve our data quality, we need to clarify the process of data collection and validation. Specifically, we require answers to the following questions:

- What methods are currently in place for validating data upon entry, especially concerning brand attribution and user details?

- Can we implement a more robust system for monitoring and cleaning data regularly to prevent these types of errors from accumulating?

Looking ahead, we anticipate potential performance and scaling concerns, especially as our datasets grow in size and complexity. To mitigate these risks, we propose the development of a scalable data architecture that can handle increased volumes efficiently. This includes optimizing our data storage practices, improving query performance, and ensuring that our data processing capabilities can scale dynamically with demand.

Your insights and feedback on how to proceed with these improvements would be invaluable. We believe that by addressing these data quality issues head-on, we can significantly enhance our analytical capabilities, enabling more accurate and actionable insights for our business strategies.

Thank you for your attention to these matters. I look forward to discussing our next steps.

Best regards,

Santosh