PART 4: Communication with Stakeholder

Subject: Addressing Data Quality and Optimization Considerations

Hello Team,

I hope this email finds you well.
As I immerse in data analysis, it's crucial to address certain data quality issues that have come to my attention. Maintaining accurate and reliable data is integral to generating meaningful insights.

Null Values in User Data:
The analysis has revealed that three columns within the users dataset contain Null values. This discovery carries implications for the validity of our findings. To ensure accuracy I took following steps: removing the Null values or substituting them with the Mean or Median. By doing so, I can prevent misinterpretations arising from incomplete data.

User_ID Duplicates in Users Table:
A critical issue has emerged within the users table, where out of the 495 entries, only 212 unique user_id values exist. This discrepancy poses a substantial data quality concern. Since user_id serves as the primary key for the users table. This inconsistency not only impacts table joins but also affects the reliability of the analysis.

Complexity of Rewards Receipt Items List:
The rewards receipt items list employs a nested dictionary structure. While this structure offers flexibility, it also presents challenges in data cleansing. A notable hurdle arises when receipt items lack an associated ID, impeding the ability to ascertain which user the receipt item belongs to.

Challenges with Purchased Date and NAT Values:
The presence of NAT (Not-a-Time) values in the Purchased Date column has been identified as a potential concern. This issue can hinder time series analysis. Ensuring consistent and valid time data is crucial to maintain the integrity of time-based analyses.

Missing Product Descriptions in Orders Data:
Within the Orders dataset, 173 products lack descriptions. This gap in information can lead to misleading interpretations and compromise the quality of the analysis.

Data Inconsistencies in Brands Table:
The Brands table highlights inconsistencies, with 54 brand codes matching barcode values and 234 instances of missing brand codes.

As we work towards comprehensive and reliable insights, your input and collaboration are invaluable. Let's collectively address these data quality issues to ensure the integrity of our analysis results. If you have insights, suggestions, or strategies to share, please don't hesitate to do so.

Thank you for your dedication to maintaining the highest standards of data quality and analysis.

Warm regards,
Swamini Bhoir