Fourth: Communicate with Stakeholders

Hello Everyone,

I hope you are doing well. As part of the Data Quality Review on the Receipts, Brand, and Users data, below are some key observations:

1. **Missing data:**

    - **Finished Date**: Around 50% of records are missing a "Finished Date." This absence makes it difficult to track the completion timelines of receipts, which may impact lifecycle analysis.

    - **Points Awarded Date**: 52% of user records lack a "Points Awarded Date," making it challenging to determine when points were granted and potentially affecting any analysis on points distribution.

    - **Points Earned**: About 45% of "Points Earned" values are missing. Additionally, some entries contain a "NaN" value instead of a zero, which may indicate either an issue with data capture or that 0 points are being recorded incorrectly. This could lead to inconsistencies in reporting and analysis.

    - **Purchased Item Count & Rewards Receipt Item List**: Both columns show a large amount of missing data, making it difficult to analyze user purchases or rewards item details. These gaps are hindering trend identification and in-depth analysis in these areas.

    - **Total Spent**: Approximately 38% of the "Total Spent" values are missing, which is a significant concern. Missing expenditure data undermines the reliability of financial analyses and could result in inaccurate insights into customer spending behavior.

    - **Top Brand:** About 52% of the data is missing

    - **Category Code:** About 55% of the category code is missing, which will lead to an issue while doing in-depth analysis on categories

    - **State, Sign Up Source:** Not all users have State and Signup Source data

2. **Duplicate Data:** During my review of the User table, I noticed that there are multiple records for the same user(_id column). I would appreciate any insights into why this might be the case, as it could help us understand how the data is being recorded and ensure its accuracy.

3. **Embedded Data:** Much of the data is stored in dictionaries (date columns like **"createDate",”rewardsReceiptItemList”, “_id”**) that require additional steps to process. Removing this could help reduce errors and help in simplifying the data, but I'd like to understand if it's necessary for how the product is set up and works.

4. I observed that the **'brandCodes'** in the dataset are currently stored as non-numerical values. I recommend using numerical foreign keys instead. This approach can help avoid issues that might arise from small differences in the string format when joining data, ensuring smoother data integration and analysis.

Thank you for your help in clarifying this matter. I recommend further investigation into the cause of the missing data, followed by appropriate data imputation or other corrective measures to improve the quality of the dataset.

Please let me know what times work for you to discuss more in detail.

Best,
Lokesh