Email to stakeholder

Hi Stakeholder,

I hope you are doing well. I have assessed the data and built a schema based on it. I performed an analysis and have a few questions.

Data Discrepancies and Questions

Questions about the data:

After reviewing the data, I wanted to understand the reasons behind several discrepancies, such as:

1. **Missing Data**:
   - **finishedDate**: Nearly half of the receipts lack this date, which is crucial for determining when a receipt becomes invalid.
   - **pointsEarned**: Many instances of this field are missing, indicating that points earned weren't consistently recorded.
   - **purchasedItemCount**: Significant gaps here could affect our ability to verify if users qualify for offers requiring multiple purchases.
   - **totalSpent and rewardsReceiptItemList**: Missing data in these fields means we also lack corresponding points earned information.
   - **topBrand and categoryCode**: Incomplete data for these fields can impact our brand highlighting and categorization efforts.
2. **Duplicate Records**:
   - More than half of the records in the Users data are duplicates, necessitating a database review to eliminate redundancies and prevent future occurrences.
3. **Inconsistent Date Formats**:
   - The date formats are inconsistent, deviating from the standard MM/DD/YYYY format, which should be standardized.

We have identified some issues through initial observation, while others were uncovered using data analytics in Python. I need information on which duplicate records should be considered correct. For records with missing data, do you recommend using default values or handling them as empty?

We have built a schema based on the sample data and have several strategies in mind.

Please feel free to ask any questions you may have. Let me know when you are available for a call so we can discuss strategies and next steps.

Thanks & Regards,
Jaswanth Reddy