Subject: Data Quality Insights and Proposed Next Steps

Hi Stakeholder,

I hope this email finds you well. During a recent review of the JSON data for users, receipts, and brands, I identified several data quality issues that could impact our ability to make accurate and actionable business decisions. I wanted to share my findings, their potential implications, and recommendations to address them.

Key Findings

1. **Receipts Data:**
    - Missing values are a significant issue, with **51% of `bonusPointsEarned`** and **39% of `totalSpent`** records incomplete.
    - There are also outliers in `totalSpent`—some exceeding **$2,000**—which require further investigation to determine if they are valid.
2. **User Data:**
    - Among 495 user records, there are **283 duplicate entries** and missing values in fields like `signUpSource` (4.3%) and `state` (5%).
    - Interestingly, over **80% of users are from Wisconsin (WI)**, which may warrant further review to understand whether this is expected or an anomaly.
3. **Brand Data:**
    - The `categoryCode` field is incomplete in **58% of records**, making it challenging to analyze brand performance by category.

Impact on Business Insights

These issues limit the accuracy of key analyses, including user behavior, transaction trends, and brand performance. Missing and incomplete data reduce our ability to draw reliable conclusions, while outliers and duplicates can distort insights. For example, understanding user engagement across states or accurately categorizing brand transactions is difficult without addressing these gaps.

Proposed Actions

To mitigate these challenges, I recommend:

- Investigating the root causes of missing and duplicate data, potentially stemming from ingestion errors or manual entry.
- Cleaning key fields such as *totalSpent* and *categoryCode* to improve the reliability of our analyses.
- Introducing **automated validation checks** during data ingestion to proactively catch issues like missing values or outliers.

Additionally, it would be helpful to align on a few points:

- Should we prioritize cleaning specific fields based on their business impact? For instance, resolving missing totalSpent or duplicate user data may significantly enhance our transaction analyses.
- Do the high totalSpent values (e.g., > $2,000) represent legitimate transactions, or should they be flagged for further review?
- Should we investigate why over 80% of users are concentrated in Wisconsin (WI), or is this aligned with our business expectations?

Next Steps

If this aligns with your priorities, I can proceed with further investigation and collaborate with relevant teams to address these issues. Let me know your thoughts or if you'd like to discuss this further in our next meeting.

Looking forward to your feedback!

Best regards,
Starry Zhang