Subject: Questions, Issues, and Next Steps regarding our Reward Datasets

Dear [product or business leader's name],

This is Julius, and I have been working on our reward datasets, specifically focusing on users, brands, and receipts. I have encountered a few questions and identified some issues that I believe are worth discussing. I would appreciate your insights and guidance on these matters.

Questions:

Relationship between barcode and brand: I would like to gain a deeper understanding of the relationship between barcodes and brands. While my understanding suggests that a brand can have multiple barcodes, our current brands dataset indicates that each brand is associated with only one barcode. To ensure we have an optimal database design, it would be helpful to clarify this relationship.

Meaning of userFlagged columns: In examining the receipts item data, I came across the userFlagged columns, such as userFlaggedBarcode and userFlaggedPrice. I believe it might be possible to consolidate these columns into other existing ones, such as barcode and finalPrice, respectively. Could you shed some light on the intended purpose and significance of these userFlagged columns?

Data Quality Issues:

Missing values: The datasets contain a significant number of missing values, particularly in the receipts dataset. Most columns have missing values ranging from 30% to 50%. Additionally, approximately 55% of items lack a barcode. Is there a way to address these missing values? If not, it is crucial to understand the root cause to mitigate any potential biases in our analysis.

Duplicate data: The users dataset exhibits a high number of duplicate rows, accounting for over half of the dataset. It is essential to identify the reason behind this issue since users are a vital component for further analysis, such as customer lifetime value predictions. Understanding the cause will enable us to rectify this problem effectively.

Data type issue: To enhance data consistency and ease of use, it is recommended to consider utilizing the datetime datatype for our date-related columns. This adjustment will streamline work processes for everyone involved.

Imbalanced data: Certain columns within the datasets exhibit an imbalance, resulting in uneven distribution. For example, the sign-up source column reveals a significant disparity, with 204 users signing up via email and only 3 users using Google. This imbalance can introduce bias in advanced analysis and predictive models. To address this concern, I propose leveraging data augmentation techniques if we move on advanced analysis and predictive models.

Next Steps:

Understanding our goals and expectations: It would greatly assist me if you could outline the business problems, goals, or expectations we are currently working towards. With this information, we can address the identified issues in a manner that aligns with our objectives and devise more effective strategies to achieve them.

Request for additional data: I believe it would be highly beneficial to compile a comprehensive database encompassing most, if not all, of our datasets. This would enable us to conduct in-depth data analysis and facilitate the development of accurate machine learning models for predictive purposes. If possible, I kindly request your consideration in providing access to more data.

Please feel free to reach out if you have any questions or require further clarification. I am eager to discuss the details with you and seek resolutions that will enhance our data quality and analysis. I look forward to your response.

Thank you for your time and attention.

Warm regards,

Julius Lee