Dear [Manager/Teamlead name],

I hope you are doing well. After going through the Fetch Rewards data file (Brands, Users and Receipt). I came across a few findings and found a few data issues hoping to have a detailed discussion.

To gain a better understanding of the relationship between barcodes and brands, it is important to note that while my current understanding suggests that a brand can have multiple barcodes, the information in our existing brands dataset indicates that each brand is associated with only one barcode. To ensure the optimal design of our database, it would be beneficial to clarify this relationship. This could involve further investigation and data validation to determine if the current dataset accurately represents the relationship between barcodes and brands.

During the examination of the receipt item data, I noticed the presence of userFlagged columns, such as userFlagged Barcode and userFlagged Price. It appears that these columns might potentially be consolidated into other existing columns, such as barcode and finalPrice, respectively. To gain a clearer understanding of the purpose and significance of these userFlagged columns, I would appreciate further clarification on their intended use and the specific meaning attributed to them in the context of our data.

Data Quality Issues:
I have identified missing values from multiple columns within the "brands" dataset(category, categoryCode, topBrand and brandCode). These gaps in data can impact our ability to accurately assess brand performance and generate meaningful metrics at different levels of granularity. To address this issue, rectifying these gaps and filling in the missing values, we can enhance our ability to precisely identify and analyze brand performance across various dimensions. The presence of class imbalance in 'category' and 'cpg_ref' should be taken into account for further analysis or model development.

In the "users" dataset, we have observed missing values in the "signUpSource", "state", and "lastLogin" columns. Additionally, more than half of the dataset has duplicate entries, which could potentially result in increased storage consumption as the dataset grows. While these missing values may not have a significant impact on the project at present, acquiring this information could enhance our analytical processes. For example, it could aid in examining user retention patterns and predicting user behaviors. By addressing the missing values and eliminating duplicates, we can ensure data integrity and leverage the full potential of the dataset for our analytical endeavors.

The "receipts" dataset exhibits missing values in various columns, and an interesting observation is the presence of 552 barcode values that are not found in the "brands" dataset. This mismatch

hinders accurate data grouping and analysis. However, if we can retrieve the missing barcode information, we can gain better insights through proper data grouping and meaningful summary statistics. Additionally, the presence of a '0' value in the "purchasedItemCount" column requires further examination to understand its significance in relation to the dataset.

Thank you for your time and attention.

Regards,
Vinay