Subject: Findings and Data Quality Issues

Hello Team,

While conducting a detailed exploratory analysis of the given data sources, I came across a few findings that I would like to shed some light on and hope to have a detailed discussion soon on the same.

- I noticed that certain data elements related to user flags, user reviews are merged with barcode, item price, and purchase count. I would appreciate some clarification on why these different aspects are combined in the same dataset, as it may affect the complexity and relevance of the analysis.
- I would like to understand the meanings of the following attributes within the list of Items in the receipts data file:
    - 'finalPrice'
    - 'itemPrice'
    - 'needsFetchReview'
    - 'partnerItemId'
    - 'preventTargetGapPoints'
    - 'quantityPurchased'
    - 'userFlaggedBarcode
    - 'userFlaggedNewItem'
    - 'userFlaggedPrice'
    - 'userFlaggedQuantity'
    - 'receipt_id'
    - 'needsFetchReviewReason'
    - 'pointsNotAwardedReason'
    - 'pointsPayerId', 'rewardsGroup'
    - 'rewardsProductPartnerId'
    - 'userFlaggedDescription'
    - 'originalMetaBriteBarcode'
    - 'originalMetaBriteDescription'
    - 'brandCode'
    - 'competitorRewardsGroup'
    - 'discountedItemPrice'
    - 'originalReceiptItemText'
    - 'itemNumber'
    - 'originalMetaBriteQuantityPurchased'
    - 'pointsEarned'
    - 'targetPrice'
    - 'competitiveProduct'
    - 'originalFinalPrice'
    - 'originalMetaBriteItemPrice'
    - 'deleted'
    - 'priceAfterCoupon'
    - 'metabriteCampaignId'

- I also observed some discrepancies between the 'category code' and 'category' fields, as well as the 'brand code' and 'name' fields. Understanding the reasons behind these inconsistencies will allow us to make more informed decisions when analyzing and reporting on the data.

The data has been modeled with the following assumptions, please clarify if these are correctly assumed or if there are any misunderstandings within these:

1. A user may exist without ever having scanned even one receipt.
2. Each user may scan one or more receipts.
3. Each receipt must contain at least one or more items.
4. Each brand may have one or more items.
5. Each item with a unique barcode must belong to one particular brand only.
6. There could be items that may not have been purchased in any receipt ever scanned.
7. Each item belonging to a particular brand may have been bought more than once by one or more users.
8. When a receipt is scanned, it triggers one or more reward events based on the number of items in the purchase.
9. Each reward event is triggered for one particular item and specifies ‘quantityPurchased’ of that item.
10. A brand must sell products in only one category.
11. Many brands could sell in the same category.

Coming to the data quality issues.

- The record shows 495 users out of which, only 212 user accounts are unique. Based on further digging, there were 283 duplicate records i.e., many users were shown logged in multiple times at the same instance. Due to this, we don’t have users' last login details which is very much required for conducting RFME analysis which helps bucketing users based on their value. Also, it was said that the role is set to a constant value “CONSUMER” but records were found to have different values.
- Approximately 50% of the values are missing for most variables in data related to rewards. Missing values primarily consist of entries without barcodes or product purchases. These entries are not associated with rewards or user flags either. This situation warrants closer examination as it may indicate potential issues, such as unrecognized data or system glitches, which need to be addressed and resolved for accurate data analysis.

I would also like to know if the company has any plans for conducting A/B testing or bucketing users into VIP categories to effectively dash out bonuses. That way we can structure our data assets in a way ensuring time series analysis can be performed easily. Moreover, moving into the future, as the company and data grow, I was wondering if there are any plans of migrating the data to a data warehouse or a data lakehouse.

I’m looking forward to our call for a more detailed discussion and analysis of the findings mentioned above.

Thanks and Regards,
Debanjali Saha