

## Fetch Rewards Coding Exercise - Data Analyst

### Communicate with Stakeholders

**Subject:** Findings and Data Quality Issues

**Body:**

Hello Team,

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

- Firstly, 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.
- Secondly, I would like to understand the significance and purpose of the 'partnerItemId' field. Additionally, I'm curious if the combination of 'partnerItemId' and 'barcode' can be used as a primary key for the dataset, as a combination of the two gives unique values hence facilitating data integrity.
- Lastly, I 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.

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 which 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.

Ajay Venkatesh Pandravada

- For columns 'pointsEarned', 'purchasedItemCount', 'totalSpent', few values are very large as compared to most values in respective fields. I would recommend investigating the outliers to ensure all processes in our App that produce these values are running smoothly which will help to determine whether the entries are legitimate or result of something erroneous that is happening.

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.

Moving into the future, as the company and data grows, I was wondering if there 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 over the findings mentioned above.

Thanks and Regards,  
Ajay Venkatesh Pandravada