

COMMUNICATION WITH BUSINESS STAKEHOLDERS

Hello Everyone,

I hope this email finds you in the best spirits. I am Bhavya Batra, I recently finished working with the data for the Fetch rewards. The three data files in JSON format were receipts, brands, and users. I was hoping to give you all an update on my progress, findings, and some data quality issues that I found.

First, I would like to summarize my approach to dealing with these data files. The data was in zipped JSON files, so I performed Data Cleaning and Preprocessing to convert and transform the data in a more structured form i.e in a data frame using Python. After this, I checked for some data quality issues. I noticed some issues within the data which I would like to discuss with you.

1. I wanted to check if the attribute `_id` is unique for the receipt, but I noticed that there are multiple records for one `_id`. I was hoping to understand that if `_id` is the primary key in the receipts table then why is it not unique?
2. The next thing I noticed was that there were a lot of unique brand codes in the receipts data that did not exist in the brand data. This could be a problem when we map them.
3. There were some records in the brand table that had the same brand code and brand name which doesn't make sense as both are different attributes.
4. Last, some of the data (rewardsReceiptItemList, dates, CPG) were embedded in dictionaries that required an extra step. It would be great if we get them expanded beforehand but I will understand if it has to be like that.

Thank you for your time and consideration. I would welcome the opportunity to discuss these points further and schedule a meeting to elaborate on these questions. Please do not hesitate to contact me.

I look forward to hearing from you.

Sincerely,
Bhavya Batra