Mail Subject : Regarding data quality issues and suggestions.

Hi Product/Business leader,

I am Data Analyst, part of data science and analytics team. I have been exploring the data namely brands, receipts and users data and I have seen some potential data quality issues. I will be sharing them with you. If you would like to add points on business process why the data is in such manner I would be delighted to know about them and understand.

The following are the potential data quality issues:

1. In the brands there are lot of missing information for “category code”. For category such as baking, beverages there seems to be already category code assigned and still I see some missing category code for baking, beverages categories.

Suggestion: I am assuming as the category is string, there might be some issue while processing the strings and assigning category code to them for example, “baking” and “baking “ are treated as two different categories though they are same due to the extra space at tail.

1. In the receipts data there is “receipt item list” column and these items have some “brand codes” assigned. The majority of these brand codes are not available in brands table which suggests that brands table information is not up to date.

Suggestion : We need to update the brands table regularly to have accurate information.

1. There are lot of duplicates in the users table. Can you tell me why there are lot of duplicate information in this table.
2. In the data dictionary provided it was mentioned in the data dictionary role is constant value set to 'CONSUMER’ but I can see there are more than one role in the data (['consumer', 'fetch-staff']).The data dictionary is not clear in this. Can you clarify me what are these different roles mean?

Suggestion : Update the data dictionary with accurate information on this.

1. In the json files I have noticed that most of columns has nested dictionary like purchase Date has {$date : value}. May I know why this is necessary? If that is not necessary then it would make the data processing one step less.
2. In the brands table there is “cpg” column and it has two different types of references information such as cpgs, cogs. It was not clear in data dictionary what does these two references mean. Can you elaborate about them?

These are some of the issues, queries and suggestions I have after exploring the data. If I have these information I could provide better visualization insights on the data which would help us. Let me know if anything was not clear from my side. Looking forward for you responses on these.

Thanks & Regards,

Gvs Goutham