Hi ______,

After careful inspection of the data provided, I have a few questions I'd like to ask.

1. Within the Receipts data schema, there is a column that contains a separate schema, which I'd call Receipts Item List schema. Inside the Receipts Item List schema, I notice that there are no fields related to each item's brand's ID. Have you collected the data on the brands' IDs when users buy items? Having brands' IDs on users' receipts would greatly help scaling the data if we want to relate items to Brands data schema.
2. A follow-up to the question above, if brands' IDs are not collected, I will be using brand codes to relate items from Receipt Item List data schema to Brand data schema. However, it seems that there are gaps in the data of brand codes in both Receipt Item List schema and Brand data schema. Are there any other data sources that could help fill in the gaps?
3. To offer better performance for the data model, I've decided to remove some columns within the Receipts Item List schema. I'd like to ask what MetaBrite is. Columns related to MetaBrite are on my list of removal as I do not see any importance of those columns, however I'd like to double-check with you before removing them.

As mentioned in question 2 above, there are gaps in the brand code in both Receipt Item List schema and Brand schema. In addition, the spelling of some brand codes are different between the two schemas, such as Ben and Jerrys vs Ben & Jerry's. The issue is discovered when I was doing exploratory data analysis on all schemas. To fix this issue, I'd need a list of correct brand codes so I can help fix the spelling issue. As for the gap, I'd require either another data source that can provide more complete data, a table that can map product item to their respective correct brand codes, or, if possible, a way to provide ID to each item's brand in the Receipt Item List schema.

Looking at the provided data, the Receipt Item List schema would be the biggest one of the 4 schemas, as it contains all items listed in each user's receipts. This could become a liability in performance when designing the data model. A method to fix this is to decrease the amount of columns, which I mentioned in question 3. Another method is to create indexes for the schema. For example, by looking up 1 receipt ID, we will use that as index to get us a table that includes the items in that receipt.

If you have any questions, feel free to ask me.

Best regards,

Nanhsun Yuan