Question 4

--What questions do you have about the data?

--How did you discover the data quality issues?

--What do you need to know to resolve the data quality issues?

--What other information would you need to help you optimize the data assets you're trying to create?

--What performance and scaling concerns do you anticipate in production and how do you plan to address them?

Hi,

Hope this email finds you well! This is Yiwei at Fetch Rewards Data Analytics Department. I am reaching out to share with you about the data quality issues we found during the analysis and potential solutions to fix the problems and to improve the situation.

Our business is to generate insights on behaviors of registered users with our platform to improve business decisions from data that includes user, brand and receipt information.

Some problems on data quality that I find through joining tables and referencing data are:

1. Null values on important metrics (ID, brands, creation date) exist extensively in the table that prevents effective data mapping and reference.
2. The interpretations for date columns are ambiguous as large integers are shown instead of timestamps. In my analysis, those numbers are interpreted as milliseconds and are subsequently converted to days. My method introduces high granularity but that only provides a relative comparison across time (i.e. time difference between instances instead of actual timestamp). However, the interpretation is subjective in order to meet analytical needs and will not be settled until we have a defined metric of the dates.
3. Primary key violation occurs for the users table as I observe duplicate user IDs.

Below are the steps to fix the data quality issues:

1. Null values for the important metrics such as ID, brands, etc. need to be populated and formatted accordingly to the existing values.
2. A more detailed data definition is necessary to interpret the date columns so that we could better understand the dates in terms of timestamps.
3. Ensure ID columns do not have duplicate values so we can reference each row specifically with its ID.

That said, there are still some information that is helpful to further interpret the data and make the best use out of it.

1. On the user side: we can include the time duration browsing our website for each user on the platform to see how long it takes them from making purchase decisions to final payments to inform our development teams and UI/UX teams to improve user experience and efficiency. This not only retains and introduces users for our platform, but also reduces system burden to serve multiple users at the same time.
2. On the brand side: we need more information regarding CPG. As each brand ID could refer to multiple CPG, the latter could provide more granular information on customer behavior and purchases to help us analyze in more details to optimize their experience and efficiency.
3. On the receipt side: data on receipts with returned items and return reasons could be documented to analyze the returns of the users after purchases. This could provide insights on improving products, operating systems and supply chains with the merchandize.

The suggestions above are the first-aids I believe could help solve the data quality problems, but we could discuss more on the further actions upon that or if you have any questions!

While composing this message, I came up with several performance concerns that might stem from our improvements.

1. Data size and complexities will compound, especially after we join multiple tables after introducing more in-depth metrics. Potential solution to this is to archive base-level data and tables in a database as a reserve and reference point for further needs of data extraction while querying specific data of interest (tables, columns, rows etc.) into data lakes/servers to analyze specific problems.
2. Separate tables into multiple tables with the primary key based on information categories for further detailed analysis (e.g. brand id and CPG id should each lead a separate table that specifically analyze brand info and CPG info as we could always join the two with brand ID.)
3. User table and brand table cannot be joined directly using respective primary keys as many-to-many relationship might be introduced and cause complications. However, after introducing a separate table from the previous step, we can join users with other metrics such as the CPG ID or other brand metrics. In this way, we can easily join users and brands for analysis without introducing the receipt table that could unnecessarily increase the data size we work with.

These are some initial steps I have thought about and please feel free to reach out for any comments or questions.

Warm regards,

Yiwei