Hello product/business leader,

I have completed the initial exploratory data analysis on the data provided in the receipts, users, and brand data files. During this analysis, I made several observations regarding data quality and have a few questions and concerns to discuss further.

In terms of data quality, after cleaning the data (flattening data, removing duplicates, transforming lists), I made the following observations:

- There are 186 additional distinct brands present on the receipts that are missing from the brands data I was provided.
- The receipts data contains 117 users that are missing from the users data.
- In the users data, over half the users were duplicates. Once duplicate values were removed, the number of users decreased from 495 to 212.
- The brands data contains many “test brands”. Of the brands in the data, 429 have “test brand” in the name.
- Receipts data is missing for December 2020 and sparse for October 2020 (2 receipts), November 2020 (6 receipts), and March 2021 (30 receipts).

As a result of these findings, I have the following questions and concerns:

- Is the sample data provided from production data? If so, can the missing brands, users, and receipts data be provided?
- The sample data was provided in JSON files, which required formatting into a usable format. As the quantity of data increases, this could result in performance and scaling issues. If the source data is already stored in a central database, then access to that database or mirror would be preferred.

I'm interested in discussing these points further and would like to schedule a meeting to elaborate on these observations. Please let me know what time works best for you.

Thanks,

Ritchie