Subject: Observations and questions about your receipts data

9/15/2021

Dear [Stakeholder],

I reviewed the receipts data you sent to our team and have a few observations and questions that I think we should address before loading the data into our data warehouse. I have also attached a proposed database schema that you can forward to your reporting team for their review; please let me know if they have any feedback, questions, or other comments on how the proposed schema will meet your business needs.

In our proposed database, we're hoping to use the "BrandCode" field from each receipt item to link receipts to brands. Through my initial exploration of the sample data with our python tools, I discovered that 62.4% of receipt lines, accounting for 67.6% of spending, do not have a brand code and therefore cannot be associated with a brand. This jeopardizes our ability to reliably answer several of your business questions because there are likely a large number of brands that we cannot summarize data for, and therefore cannot say with confidence which brands have the most receipts scanned, highest spend, or most transactions. Will it be possible in the final dataset to ensure the BrandCode field is more complete so that queries intended for your questions return accurate results? If not, what are your thoughts on the impact of this data issue on your business needs?

I also observed that individual receipt lines have up to three different price data points: the item price, the final price, and the user flagged price. They are typically all equal, but don't always all have values. I also noticed the receipts have a total spent value; however, in a few anecdotal cases where receipt lines only contained a user flagged price, the sum of prices on the receipt did not equal the marked total price. Which price values should we consider the source of truth for the amount a user spent on an individual item or on a receipt overall?

These first two questions are, I believe, the most important to address for the success of this project. I also have a few others that are worth looking into:

- Is the "SignUpSource" of individual users relevant for your business needs? It's not in the schema you provided, but it is in the Users dataset.
- How can we reliably identify and remove test brands from the brands dataset?
- What should be done with receipt lines with a barcode of 4011 (Item Not Found)?
- Is it possible, at some point in the future, data will be added to the Users dataset with a role other than "consumer"?

From a database perspective, I don't have any serious concerns about the scalability of our data model for these data, though I plan to create an index in the database to improve efficiency of queries that use both receipt and brand data.

Please let me know when you have any information that could help us address the questions above, as well as if you have any follow-up questions!

Thanks,
David Marchena
Data Company data analyst