To whomsoever it may concern,

I’m writing this email to address a few things and ask some questions for better understanding of the data and what is the expected desired outcome to be able to produce accurate data insights and analysis.

I had a couple of questions with respect to the data in the brands table. There’s two columns bar code and brand code. Sometimes the brand code is numeric and sometimes it’s string and sometimes the value of brand code is the same as the value of bar code. So, what’s the significance of the two columns? Also, what does cpg imply in the brands table?

I discovered the data quality issues by loading the tables into R and querying it.

Upon looking at the data more closely and querying it, I can see that the brands table contains a mix of both production and testing data. Ideally, production data should not have any testing/staging data, and these two will live in separate tables/databases. In addition, the brand code column does not seem to have a consistent data type, as some records contain a numeric value whereas other records contain a string value.

What type of metrics are we looking to capture from this data? There may be a need to create an "items" dimension if, in the future, we want to track metrics like "top item ordered", "average spend for a specific item", etc.

What are the metrics we are concerned with the most in the receipts table? The item list is a JSON object which, as the table scales upwards in size, will be computationally expensive to unpack in SQL. Instead, if we are only concerned with a few metrics from the many fields in the item list JSON object, we can unpack those fields to separate columns in the table, which will optimize read performance.

If there are any questions or you need more elaboration on anything, please reach out to me via this email ID.

Looking forward to hearing back!

Thanks for your time and assistance.

Regards,

Akhila