Communicate with Stakeholders

Upon analysis of the 3 data files provided, I wanted to share some findings and questions and I would love to hear your thoughts about the same.

I see some valuable information that is retrieved from the 3 files and I have designed a schema adhering to database constraints and normalization standards to organize and store the information suitably. To do this more efficiently, I wanted to understand the value of information in the attributes like – CPG, ReceiptsItemList and repetitive columns like itemPrice, finalPrice, discountedItemPrice, originalFinalPrice.

One of the major data quality issues I see is with one of the attributes of Receipt called ReceiptItemList that in turn consists of 35 columns, many of which have more than 90% of null values. Although the 35 attributes can be sub-categorized into smaller tables, it is essential to remove these nulls before we do so. It is critical to understand the business use case or value of each of these columns, based on which we can take measures like eliminating the columns, imputing with suitable values. Additionally, the connecting attribute between Receipt Items and Brands needs evaluation to ensure it is distinct and not null.

It would also help to know the most frequent or primary information that will be retrieved using this database design, for example, if Receipts are going to be analyzed the most, many computations can be simplified and optimized by using standard procedures and indices.

I am concerned about the processing and scaling of data in the production environment due to a large number of records and multiple connections to other tables in the data that require traversal through the entire table for retrieval of information. This can be avoided by adding indices that scan through a smaller set of relevant information and can speed up the process.