

Fourth: Communicate with Stakeholders

For this question, we are writing an email to the team leader, informing her about the data quality issues we found.

Subject: Data quality issues with Receipts, Users and Brands data.

(Body of the email):

Hello team leader,

After conducting a thorough Exploratory Analysis of the records in Receipts, Users and Brands, we came across the following data quality issues that we believe are important for you to know.

1. A considerable amount of data is missing for certain fields like-
 1. finishedDate- for 49%(almost half) of the receipts we do not know when they become invalid(assuming that the date on which a receipt finishes processing is the date on which it becomes invalid)
 2. pointsEarned- 45% of the values for the 'pointsEarned' field are missing. This means that points were earned for certain receipts, but the data was not captured and that is why the large number of missing values.
 3. purchasedItemCount- large number of missing values will pose problems for deciding if users who bought more than one unit of a product qualify for special offers/bonus points that require them to purchase certain number of products/brands.
 4. totalSpent, rewardsReceiptItemList- Since data for the total amount spent on a receipt, and items shopped in a transaction is missing, it is natural that we do not have information about points earned(pointsEarned field) for those transactions.
 5. topBrand (Boolean indicator for whether the brand should be featured as a 'top brand')
 6. categoryCode (The category code that references a category of a brand)
2. For columns 'pointsEarned', 'purchasedItemCount', 'totalSpent', there are a significant number of values that seem out of place (very large as compared to most values in respective fields). We would recommend investigating the processes in our App that produce these values, to determine if they are legit or result of something erroneous that is happening.
3. There are a lot of (more than half) duplicate records in the Users data. I strongly suggest going over our database, to eliminate redundant records and ensure there is no way such anomalies can happen again.

4. Lastly, we found the date formats to be inconsistent, against the usual MM/DD/YYYY or similar standard date formats. For this too, we'd consider going over through our database to ensure that date fields are being captured and stored in a consistent manner.

We have a plan to deal with the missing values issue and other problems as well, that we'd like to discuss with you in detail. Let us know what time works best for you so that we can set up a meeting.

Best regards,
Eli Bohara
Peiting Wu
Yuhui Chen
Kengshao Yeh