

Subject: Data Quality Issues with Users, Receipts, and Brand Data

Hello [Business Leader],

I was reviewing our Users, Receipts and Brand data and discovered some data quality issues that may impact the business. I would like to discuss these concerns with you and gain some insights into the data.

Data Quality Issues:

The major issue revolves around duplicated records and missing data. More than 50% of the users data consists of duplicated records and approximately 40-60% of data is missing for the following fields: purchased date, receipt items, total spent, top brand, brand code, barcode, category, and category code. There's other missing information as well; however I wanted to highlight these fields as they are important for analysis purposes, marketing efforts, and the organizational structure of our data.

Another issue is regarding data inconsistency in the receipts and brand data. There are multiple records for some barcodes, brand names, and brand codes with varying information in the other fields. This leads to an incorrect representation of data for these brands and items. It can cause cascading issues from inconsistent models to missed marketing opportunities due to inaccurate brand representation.

In addition, I also noticed a potential exploit with the points earned for some purchases. Our records show a moderate number of receipts where there is an inconsistently large number of points earned for very little total spent amount. If this is an exploit of the points calculation system, it is crucial to patch as soon as possible.

Addressing the issues:

We need to improve our data collection process and enforce mandatory fields on data entry. This will decrease the amount of important information that is missing from the data, like total spent and purchased date. Our data transformation and loading processes also need to be refined to address the data inconsistency issues. We can address the duplicate and inconsistent records in this step and also ensure fields like brand code and category code are properly coded into all the tables. A validation process should be created to check the data periodically and ensure data quality. Lastly, we need to inspect the point calculation system to fix the potential exploit.

In order to address these issues, can you provide me with the following information?

- What are the data sources and data collection procedures?
- Are there any known issues that are causing missing or inconsistent data?
- What information is crucial to include for the data entry process?
- What is the logic behind the points calculation system?

Regards,
Vedhik Kodavatiganti