Email to Stakeholder

Hello Stakeholder,

I am the Data Analyst assigned to work with data in the Receipts, Users, and Brands files. After conducting my initial analysis of this data, I found some major data quality issues.

- The first issue I noticed was that the JSON files provided to me for analysis are not in the correct JSON file format. I managed to parse these files and extract the necessary data using python. To speed up the analysis process, I suggest communicating with the source of this data and figuring out a solution to organize the data in the correct JSON file format.
- I found the date to be in an unusual format. This is likely to be an encoding issue; to resolve this, I would need to know the encoding used by the source of these files.
- When analyzing the Receipts file, we can see that the rewards receipt item list data field contains another table within it. I separated this field into a new table called transactions for sake of normalization and to make it suitable for a relational data model.
- I have some questions regarding this new table though. It has 47 columns, which is odd. Many of these columns have over 7000 of their 7382 records nulls. Further analyzing the raw data in the file (before parsing), I found these columns to be inconsistent across the records with some columns existing for one record, but not the other. It would be helpful to know which columns are important and which ones can be left out.

- I propose using the combination of Receipt_Id and barcode as the primary key for this new transaction table. I would like to confirm if this would retain uniqueness for each row and be suitable for being the primary key.
- The data is null and duplicate values
    - The Users table has 283 duplicated ids and null values in the columns for Last Login, sign-up score, and state
    - The Brands table has 936 duplicated ids, which are about 82% of the total records it has. The columns of category, category code, top brand, and brand code have a significant amount of null values
    - The Receipts table has no duplicates, but more than half of the columns are around 40% to 50% of their null values.

I believe it would be best to set up a meeting so I could walk you through my analysis, and clarify questions.

Best,

Abhishek