Hi Team,

I hope this email finds you well. I wanted to share key findings from the recent data quality analysis across the **Users**, **Products**,and **Transactions** tables, and highlight an interesting trend in the data. Additionally, I have some outstanding questions that will require further clarification or action.

Key Data Quality Issues:

* Missing Values: Several columns included missing values in all the tables. The fields such as birth\_date, gender, or other categorial variables including missing values may impact the accuracy of analyzing the products. Key column of Barcode also included missing values, which is a data quality issue that must be prevented.
* Duplicate Rows: There were duplicate rows in Products and Transactions data, which were cleaned to ensure accurate results. The data quality analysis included steps of cleaning the data to remove the duplicates and convert the unstructured data to a structured table for analysis. Initially, the quantity and sales columns contained duplicates, missing, or inaccurate formats, so it was cleaned to retain accurate records. There were duplicate created dates values with different birth date, which may be missing values, and can indicate multiple accounts on specific dates.
* Data Format: The Barcode and Date columns had inconsistent format, which were cleaned to standardize so they can be used between the tables accurately.
* Data Distribution: Descriptive statistitcs and value counts were collected for all columns giving insights on the range or common categories. It showed the most popular Brands (Rem Brand, Private Label; Brand not known, which is a data quality issue for missing Brand name); Manufacturer (Placeholder Manufacturer, Rem Manufacturer, etc.), and Categories (Health & Wellness, Snacks, Hair Care, Medicine & Treatment, etc., although the Category columns were similar to each other so clustering can be performed to classify the categories better.

Interesting Trend:

* Top Brands by Receipts: Among users aged 21 and over, the top 5 brands by the number of receipts scanned are Nerdscandy, Dove, Great Value, Sour Patch Kids, and Trident. This indicates strong brand loyalty or popularity among adult users.

Request for Action:

* Data Cleaning: Addressing the missing values, duplicates, and data format in the datasets is essential. Implementing a data cleaning routine to handle these issues systematically is recommended.
* Field Clarification: Some fields, such as category\_3 and category\_4 in the Products dataset, are not well-defined. Getting more context on the fields can help clean or transform the raw source data more accurately to be analyzed correctly.

Thank you for your attention. Please let me know if you need further details or have any questions.

Best regards,

Ankur