Third: communicate with stakeholders

Considering Tim as the Director for the product, this mail is constructed to give a summary for Summary of Data Quality Issues, Trends, and Required Actions.

Subject: Summary of Data Quality Issues, Trends, and Required Actions

Hi Tim,

I hope you're doing well. I'd like to provide a concise update on the key data quality issues and outstanding questions we've encountered. Additionally, I've highlighted an interesting trend in the data and outlined the support needed to address any unresolved concerns.

Key Trend:

One important trend in the data is that snacks **are the most popular purchase across all generations**, from Baby Boomers to Generation Z. While the amount spent varies, snacks are consistently bought more than any other type of product.
This finding suggests that:

- Promoting snack deals could encourage more users to shop and spend more.
- Partnering with popular snack brands could boost sales and increase customer loyalty.

By understanding why snacks are so popular—whether it's due to pricing, brand loyalty, or promotions, Fetch can improve its strategy and encourage even more purchases.

Attached below are the **key data quality issues** observed across datasets:

USER Dataset:
1) GENDER Column: Inconsistent formatting due to:

- Case sensitivity ('non_binary' vs. 'Non-Binary').
- Use of underscores instead of spaces ('prefer_not_to_say' vs. 'Prefer not to say').
- Missing values and inconsistent representation ('unknown' vs. 'not_specified', 'not_listed' vs. 'My gender isn't listed').

2) LANGUAGE Column: 30% of data is missing.

PRODUCT Dataset:

1) Missing Data: CATEGORY_4 (92%), MANUFACTURER (26%), and BRAND (26%) columns contain significant gaps.
2) Duplicate Records: 215 duplicate rows identified.

3) BARCODE Column: Stored as a float type, making it difficult to use as a categorical variable. It also contains null values, which must be handled to enable its use as a primary key.

Challenge:
The missing data in BRAND and MANUFACTURER columns significantly reduces their effectiveness in generating meaningful insights.

TRANSACTIONS Dataset:

1) BARCODE Column: Contains null values and is stored as a float type.
2) Duplicate Rows: 215 duplicate rows identified.
3) Receipt-Level Duplication: Some rows only differ by FINAL_QUANTITY and FINAL_SALE values.

Challenges:

1) Each RECEIPT_ID has multiple entries differing only in quantity and sales values. Clarification is needed on which rows should be used for analysis.
2) The dataset lacks a primary key. A proposed approach is to split it into two datasets:
    - (Dataset 1): RECEIPT_ID, PURCHASE_DATE, SCAN_DATE, STORE_NAME, USER_ID
    - (Dataset 2): RECEIPT_ID, BARCODE, FINAL_QUANTITY, FINAL_SALE
3) Excluding rows with missing FINAL_SALE values may impact FINAL_QUANTITY accuracy.

Request for Action:
To proceed with resolving these issues and making the dataset more reliable, we require:

- Clarification on handling duplicate rows in the TRANSACTIONS dataset which only differ on FINAL_QUANTITY and FINAL_SALE values.
- Confirmation on which representation of GENDER values should be standardized.
- Guidance on how to treat missing values in key fields such as LANGUAGE, CATEGORY_4, BRAND, and MANUFACTURER.
- Direction on whether the proposed dataset restructuring aligns with the intended use of the TRANSACTIONS dataset.

Please let us know how you’d like to proceed. Looking forward to your feedback.

Thanks & Regards,
Kasi Vishwanth