Subject: Summary of Data Quality Issues and Next Steps for Analysis

Hi [Team/Product Leader],

I’ve been investigating three datasets (USER_TAKEHOME, PRODUCT_TAKEHOME, and TRANSACTION_TAKEHOME) and identified some key data quality issues and trends, as well as areas where additional clarity is needed.

USER_TAKEHOME

- **Data Quality Issues:** Found NULL values in BIRTH_DATE, STATE, LANGUAGE, and GENDER column. Also, found placeholder values like '01/01/1900', '01/01/1970' for 1282 records in BIRTH_DATE column.
- **Fields challenging to understand:** Not clear what values like “en” and "es-419" stand for in LANGUAGE column.
- **Request for action:** Clarify how to handle placeholder dates in data, convert "en" and "es-419" to "English" and "Latin American Spanish" for better clarity, and determine a strategy for managing NULL values effectively.

PRODUCT_TAKEHOME

- **Data Quality Issues:** NULL values were found across all category columns, with extremely high number in category 3 and 4. MANUFACTURER, BRAND and BARCODE contained NULL values too. BARCODE does not have fixed length of characters. A major issue is the presence of 215 duplicate records.
- **Fields challenging to understand:** Found "PLACEHOLDER MANUFACTURER" and "NONE" values in MANUFACTURER column and values like “BRAND NOT KNOWN” and “BRAND NEEDS REVIEW” in BRAND column.
- **Request for action:** Address the high number of NULL values in the Category hierarchy. Barcode lengths must be standardized to ensure consistency. Additionally, clarification is needed on how to handle placeholders and unknown values in the Manufacturer and Brand columns. Lastly, we should investigate and identify the root cause of duplicate records in the data and delete all duplicate records in current data.

TRANSACTION_TAKEHOME

- **Data Quality Issues:** NULL values found in PURCHASE_DATE, SCAN_DATE. The BARCODE does not have fixed character length and contain NULLs. FINAL_QUANTITY column had ‘zero’ values and FINAL_SALE had missing data, both these entries were duplicate records which was almost 50% of the total records. Also, found 47 records where the SCAN_DATE was before PURCHASE_DATE.
- **Fields challenging to understand:** A lot of USER_ID in TRANSACTION_TAKEHOME are not present in USER_TAKEHOME. Similarly, a lot of BARCODES in TRANSACTION_TAKEHOME are not present in PRODUCT_TAKEHOME
- **Request for action:** Handle NULL values effectively across the dataset. Barcode lengths should be standardized to align with the expected range of 12-13 digits. Investigate the root cause of 'zero' values in the FINAL_QUANTITY column to determine their validity. Additionally, discuss and decide on the appropriate action for records where SCAN_DATE occurs before PURCHASE_DATE, as this may indicate data entry errors or other inconsistencies.

One interesting trend in the data:

I have identified spending distribution based on Age Groups, State and Stores and the results are as below:
Age Groups: People aged **35-44** are the highest spenders **($167.82),** followed by **65-74 ($159.35)**, followed by **55-64 ($106.47)**.
State: People in **PA** are the highest spenders **($96.09)**, followed by **Florida ($87.36)**, followed by **New York ($50.04)**.
Store: People spend the most in **Walmart ($225.01)**, followed by **CVS ($86.83)**, followed by **SAM’s CLUB ($42.34).**

I believe addressing these data quality issues and gaining clarity on the outstanding questions will significantly improve the reliability of our analysis and help uncover further actionable insights. I’d appreciate your guidance on the requests mentioned above, and I’m happy to collaborate further to resolve these issues effectively.

Please let me know if you’d like to discuss this in more detail or schedule a quick meeting to prioritize the next steps.

Thank you for your support!

Best regards,
Rohan Shah