Third: Email communication with Stakeholders

I am writing an email to stakeholders team to explain the Data quality issues, outstanding questions, interesting trends and asking for guidance on next steps.

Email:

Subject: Data Quality Issues & Outstanding question from Exploratory Data Analysis

Hello Team,

I have conducted exploratory data analysis for User, Product, and Transaction datasets. I wanted to highlight key data quality issues and outstanding questions from these datasets.

1. Data Quality Issues:
    - Missing Data: I noticed a high percentage of data is missing in all the datasets. For example, BRAND and MANUFACTURE columns have more than 200k records missing.
    - Placeholder Values: On top of having nulls, the MANUFACTURER column contains numerous instances of "PLACEHOLDER MANUFACTURER" which would highly impact any further analysis conducted on this dataset.
    - Inconsistent Dates: There are impossible scenarios in Transaction dataset where PURCHASE_DATE is greater than SCAN_DATE. Around 100 records are impacted by this issue
    - Barcode Issue1: There are around 4000 BARCODES that have a length less than 10 raising concerns about their validity
    - Barcode Issue2: More than 20,000 transactions reference barcodes that do not exist in the Product dataset, indicating possible missing product records or data mismatches
2. Outstanding Questions:
    - Handling Outliers: There are some extreme outliers in FINAL_QUANTITY and FINAL_SALE columns. Considering the nature of these columns do you recommend dropping these outliers?
    - Special Characters: Some of the brands and store names have special characters. Is this something anticipated? Or do you recommend dropping these brand names?
    - Placeholders: How do you want us to handle the “Placeholder Manufacturer “value in Manufacture column.
    - Barcode validity: Can I consider BARCODES with length less than 10 as invalid?
    - Multiple Manufactures: Is it possible to have multiple manufacturers for the same brand?

3. Interesting Trend: Our analysis shows that users who have had their accounts for at least six months contributed the highest total sales, indicating strong retention and spending habits among long-term users. Additionally, there is no strong correlation between the number of receipts scanned and total spending
4. To ensure data accuracy, we need guidance on handling missing and placeholder values in the MANUFACTURER and BRAND columns, as well as clarifying whether barcodes with fewer than 10 digits should be considered invalid. Additionally, over 20,000 transactions reference barcodes missing from the Product dataset—is there a mapping available to resolve this? Should extreme values in FINAL_QUANTITY and FINAL_SALE be removed or treated as valid? Lastly, do some brand and store names contain special characters—do these require standardization?

Any additional business rules, documentation, or data mappings you can provide would be extremely helpful in resolving these outstanding issues. Looking forward to your input!

Best,

Akhil