Third: communicate with stakeholders

Subject: Assistance Required on Data Quality Issues, Insights on the dataset

Hello [Name],

Hope you are doing well. As part of our data analysis that was performed, I have identified several key data quality issues, I have some outstanding questions regarding the dataset as well. Additionally, I have observed a few interesting trends worth highlighting.

Key Data Quality Issues:

1. **Data Format Issues**: Data types were inconsistently assigned across tables, with most columns classified as `object`. Proper data types needs to be assigned for accurate analysis.

2. **Missing Values**: Transactions Table: Barcode column has missing data for 5,762 receipts. Users Table: Missing data in columns BIRTH_DATE - 3675, STATE - 4812, LANGUAGE - 30508, and GENDER -5892 rows. Product Table: All columns have high missing values.

3. **Data Inconsistencies:** Leading spaces in the FINAL_SALE column for 12,500 rows. FINAL_QUANTITY column includes strings such as "zero," creating inconsistencies. GENDER column has duplicate and inconsistent entries (e.g: “not_listed” , “My gender isn’t listed”) they mean the but appearing has different entries.

4. **Duplicate Rows:** Duplicates found in the Product and Transactions tables.

5. **Table relationship Issues**: The user_id in user table is expected to have all entries of the users from transaction table. But some USER_IDs in the Transactions table are not present in the Users table. Similarly, barcodes in product table is expected to have all entries of the barcodes from transaction table, but some Barcodes in the Transactions table are missing from the Product table.

Outstanding Questions:

1. CATEGORY_4 Column (Product Table): This column has a high volume of missing data and seems redundant with CATEGORY_3. Can this column be dropped if it is irrelevant, would like to know more on the purpose of the column?

2. FINAL_SALE and FINAL_QUANTITY Columns: What logic determines FINAL_SALE values, like when FINAL_QUANTITY is 0 or other scenario? Why FINAL_SALE has null values even when FINAL_QUANTITY value exists ?

3. BIRTH_DATE Null Values (Users Table): How should these null values be handled? Should I assign placeholders, or omit them?

4.Barcode Column: There are `-1` values in the Barcode column. Are these valid or should they be treated as inconsistent?

Interesting Trend:

- Female users are the top contributors to fetch app activity, with the highest engagement rates.
- Top-ranking states by user activity are Texas, Florida, and California.
- Receipts are scanned most frequently for Walmart, Dollar General Store, ALDI, Target, and Kroger.
- The top brands as per number of product exist are REM BRAND, PRIVATE LABEL, CVS, and SEGO.

Request for Action: Primary Key Assignment: There is no primary key for the Product and Transactions tables. Unique identifiers for each record need to be added.

I have handled most of the data quality issues for the dataset, but your input on these issues will be quite helpful to ensure accurate analysis and actionable insights. Please let me know if additional details are required.

Looking forward to hearing from you.

Best regards,
Purva Gharat