Email Message

Subject: Data Investigation Summary and Request for Action

Hi Team,

I wanted to share an update on the investigation we have been working on regarding the data quality and insights on the June 2024 to September 2024 data provided. Here's a high-level summary:

Key Data Quality Issues:

Missing Data: All three datasets under investigation are missing data. The User table (5 variables: missing between 4% and 30%), Products table (7 variables: missing on average about 22%), and the Transactions table (2 variables: missing 11% and 25%), affect our ability to draw accurate conclusions.

Inconsistent Formatting: Some records have inconsistent gender categories (for example, both options 'My gender isn't listed' vs not_listed present), which could skew analysis, especially when trying group data properly.

Duplicate Entries: We discovered that both the Products table and Transactions table have duplicate records, 57 records and 171 records, respectively. This duplication can lead to inflated metrics in the future.

Redundant Data & Data Anomalies: The Transaction table contained close to half of redundant records where transactions were being counted twice because for almost every sale and quantity transaction combination, a similar record was present with the only difference that its corresponding quantity was set to 0. Among other data anomalies found include:

- User age as low as 2 years and as high as 125 years.
- Scanned dates before purchase dates (94 records)
- Misspelling and Inconsistent capitalization (Products and Transactions Table).
- Decimal values for the variable "quantity", which should be a whole number.
- The word 'zero' was present in the quantity variable, instead of the value 0.

Outstanding Questions:

Source of the Missing, Inconsistent, Duplicate and Redundant Data: We need clarity on whether this data gap is due to a system error, a delay in the data pipeline, or something else.

Interesting Trends and Observations:

Interesting trends that have been observed despite the data issues include:

- A large portion of our users are in their 20s and 40s.
- Females account to close to 69% of users, while males are in the 28% range. The rest of the users account for less than 2% in other gender categories.
- 30% of our users registered in 4 states: TX (9.48%), FL (9.37%), CA (9.02%), and NY (5.99%) and this percent increases to over 50% from 9 out of the 50 states and territories.
- over 60% of products are scanned within 2 days.
- the average **account age** is about 37 months.
- In the June 2024 to September 2024 period, the type brand by sales is CVS, while 54% of the purchases in the 'Health & Wellness' category were from of Baby Boomers.
- Finally, purchases peaked in July 2024 (37%) and declined all the way to September 2024, the latter being the lowest of the four months, with only about 6%.

Request for Action:

To move forward, I would need your help in:

1. **Clarifying the cause of the missing, duplicate, and inconsisten in data** – Can we confirm whether it's a known issue or something that needs to be addressed by one of our teams?
2. **Decision on data anomaly checks** – Should we implement additional quality checks in our system?

Looking forward to your feedback!

Thanks!

Wilson