Hello Team,

We wanted to share the findings from our recent data analysis, highlighting key data quality issues, outstanding questions, and an interesting trend analysis on our customer base. Along these, I have also included some action points that may require further discussion.

Please find below data quality concerns, questions and actions pertaining to the three feeds.

1. The **Users** table has missing values in key columns such as birthdate, which hampers our ability for market identification, growth analysis, and product recommendations. The percentage of missing values has significantly increased over the last two years for birthdate (13%), gender (13%), and state (8%). However, missing values for language have decreased to 1.5%, but is above 50% historically.
   Are these fields optional in the feed for the customer? If so, can we reinforce validation to populate these in older records and moving forward?

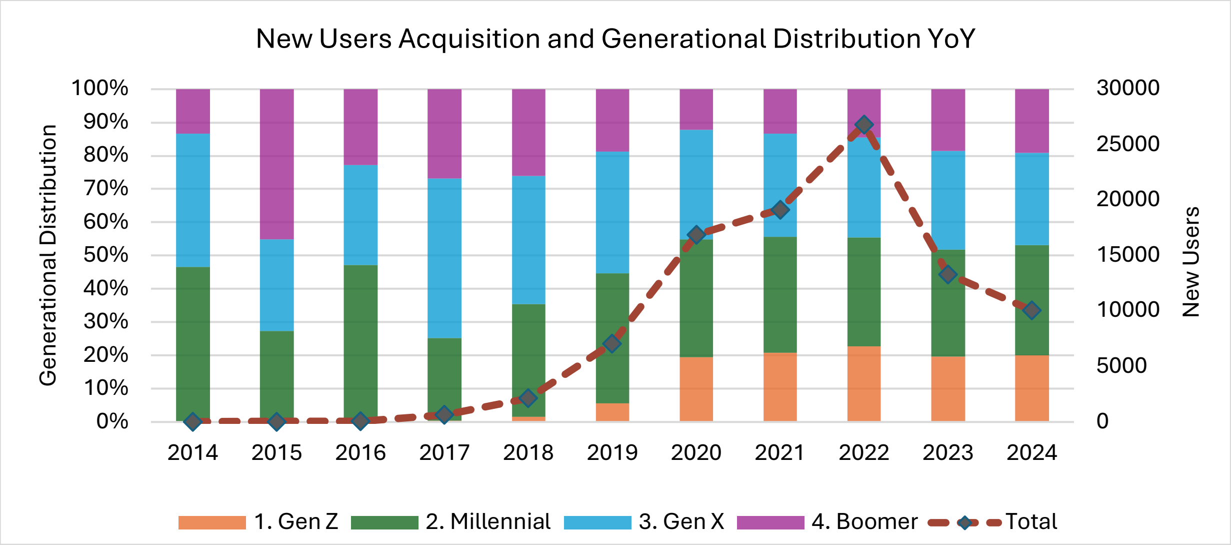
2. The **Product** table lacks a unique product identification key and a date variable to track when a product was added or modified. Some products have null values in the barcode column (<0.5%), preventing them from being linked to transactions. Additionally, barcodes overlap across different products, and there are duplicate entries for the same product (0.02% in total).
   Are these data population issues, or pertain to some logic? Can we revisit our data design to incorporate a Primary Key and a type 2 SCD (to track historic changes)?

3. The **Transactions** table has an average of ~11% missing barcodes daily, with a sharp rise observed in July (~18%) that stabilized afterwards. The missing values make it impossible to map these transactions to the Products table. It also contains a significant number of duplicate entries (~44%), with ~22% of duplicates coming from *final_quantity* column values marked as 'zero' and the other ~22% having blank values in the *final_sale* column. These two cases do not overlap—*final_quantity* zero never corresponds with a blank *final_sale*. In total, these two cases account for 50% of the data. There are also some edge cases of final_sale = 0 (<1%).
   Are all these cases considered valid transactions? If so, what is the explanation regarding them, do they cover a user journey? Additionally, for the missing barcodes, is there an alternative way to link these transactions to products and any reason for July to have a spike in missing values?

4. **Join Percentages:**
   - The join rate between the **Transactions** and **Users** tables is extremely low (0.5%), meaning only 0.5% of transactions can be mapped to a user via *user_ID*. These *user_ID* values are 24-character hexadecimal strings, likely generated through an anonymization process. Can we verify whether this process considers factors like leading zeros or whitespaces to ensure consistency across the tables?
   - The join rate between **Transactions** and **Products** tables after removing duplicates and missing barcode records is 56%.
     Is the product table missing entries of products or are these a specific subset showing only Fetch Partner products?

5. For the Users and Products tables can we request for a historic snapshot if available, that would show how their attributes changed over time to ensure accurate joining with the transactions table.

As mentioned, the team also conducted an analysis into user acquisitions YOY looking at generational distribution. Data completeness in the user table pertaining to birthdates was taken as an assumption.



We see here that early adopters were primarily Millennials and Gen X, followed by a surge in Boomers the next year. Starting in 2018, Gen Z began to gain traction, capturing and maintaining around 20% of the customer base during COVID—mostly at Boomers' expense. After COVID, our overall growth has declined. Key takeaways from this is a need to re-engage with boomers emphasizing on simplicity of the app design, create more incentives for Gen Z through social and gaming rewards and further investigate factors influencing decline in growth post pandemic.

In summary, to ensure we can address the data challenges effectively and continue with similar analysis, we'd appreciate any insights regarding the outstanding questions, particularly around data validation, join logic and historical snapshots. This would help us refine our approach and implement necessary improvements to move forward.

Please let me know if any additional details are needed or if any of these points require further discussion; I'd be happy to set up a call.

Looking forward to your thoughts.

Regards,

Muhammad Ibrahim