Respected Sir/Madam,

I wanted to share key findings from our recent data exploration of the **Users, Products, and Transactions** datasets. This review helped identify several **data quality issues** that may impact analysis and decision-making, along with some **outstanding questions** that need further clarification.

Key Data Quality Issues & Outstanding Questions:

In the **Users Data**, we identified that the values for multiple attributes such as 'Birth Date', 'State', 'language', and, 'Gender' are **missing**. One record has an account creation date **before** date of birth, which is not logical. This could be due to incorrect data entry, system errors, or a misunderstanding of how dates are recorded. One key question that remains is whether these missing values were not provided by some users or was it an input error?

In the **Products dataset**, we found that **0.48%** of product records have duplicate or missing barcodes, even though barcodes should be unique identifiers. This issue raises concerns about product tracking and could lead to difficulties in linking products to transactions. Additionally, we observed a hierarchical dependency in category fields where, if CATEGORY_1 is missing, all subsequent category fields are also missing. This dependency may indicate an expected structure, but it also creates gaps in data where incomplete records might reduce analytical accuracy. One key question is whether these missing barcodes are expected for certain product types, or if they represent gaps in data entry or errors in barcode assignment. Understanding how barcodes are assigned and managed would help determine whether these issues require corrective action.

The **Transactions dataset** has several data quality issues. **5,735** transactions are missing barcodes, making it hard to track purchased products. We also found **12,332** transactions with a sale amount but no quantity and **12,800** with a quantity but no sale amount, which could be due to weight-based pricing, discounts, or data entry errors. Additionally, **94 transactions** have a scan date before the purchase date, suggesting possible system or human errors. Further, only **261 t**ransactions have a valid user ID, **144** have a valid product barcode, and just **72** transactions have both a non-zero sale and quantity, highlighting gaps in data accuracy.

Interesting Trend in the Data:

One notable insight from our analysis is the **year-over-year customer growth pattern**. The data shows **steady growth until 2018**, followed by a **period of rapid expansion from 2018 to 2022**. Customer acquisition peaked in 2022 with **26,807 new customers**, representing the highest point of growth. However, starting in 2023, there has been a **decline in growth**, with customer additions dropping to **15,464 in 2023** and further down to **11,631 in 2024**.

This trend raises important questions about the factors influencing the decline - whether it is due to market saturation, changes in business strategy, external economic conditions, or shifting customer preferences. Further investigation is required to determine the underlying causes and identify potential strategies to sustain or reinvigorate growth.

Next Steps & Request for Action:

To resolve these outstanding issues, we need additional context and support:

- **Product & Business Input:** Are barcode duplicates and missing values expected, or do they indicate data integrity problems?
- **Data Team Assistance:** Can we investigate why SCAN_DATE is sometimes earlier than PURCHASE_DATE?
- **Policy Clarification:** Do business rules allow transactions to have a price but no quantity (or vice versa), or do we need adjustments in data processing?

I appreciate your guidance on these points and would be happy to discuss this further. Please let me know a good time to connect.

Best,
Anusha Raju
Data Analyst