Hello Stakeholder,

I wanted to share some key findings and the next steps based on the data quality review. Here's a high-level summary:

Key Questions About the Data:

1. What are the expected standards for mandatory fields in the user and transaction datasets? For example, should fields like `state`, `role`, and `signUpSource` always be populated?
2. Are there specific business rules for handling future dates or invalid timestamps in fields like `createdDate` and `lastLogin`?
3. Should duplicate records (e.g., same `user_id` or `barcode`) be resolved or flagged for further investigation?

How I Discovered Data Quality Issues:

I conducted a thorough analysis of the datasets, leveraging statistical and visualization techniques. Here are the significant issues that i identified:

- **Missing Data:** Several key columns, such as `role` and `state`, have missing values across a portion of the records. These gaps could impact segmentation and reporting accuracy.
- **Duplicate Entries:** User and brand datasets include duplicates, which could lead to inflated counts and inaccurate insights.
- **Inconsistent Formats:** Fields like `categoryCode` and `barcode` exhibit format inconsistencies, making them difficult to use reliably in analyses.

What I Need to Resolve These Issues:

Business Context:

- Clarify which fields are mandatory and what levels of missing or invalid data are acceptable for production use.

Data Source Knowledge:

- Understand how data is ingested and stored (e.g., APIs, manual uploads) to pinpoint where errors are introduced.

Handling Duplicates:

- Confirm whether duplicates should be removed, flagged, or handled as-is for further validation.

Additional Information Needed:

- **Data Lineage:** A clearer understanding of the end-to-end flow of data from source to analysis. This will help identify weak links in the pipeline.
- **Expected Volume and Frequency:** Insights into how frequently new data is ingested and the expected growth over time. This will influence our scaling approach.
- **Usage Scenarios:** Specific business questions or KPIs that this data should support, ensuring our models are aligned with key objectives.

Anticipated Performance and Scaling Concerns:

Data Volume Growth:

- As the datasets grow, slower queries and performance bottlenecks may arise. To address this, I plan to implement indexing and partitioning strategies in production.

Automated Data Cleaning:

- Introducing validation pipelines to flag or resolve issues (e.g., duplicates, invalid dates) may increase processing time. Prototyping will help balance accuracy and efficiency.
- **Duplicate Management:** Ensuring duplicates are flagged or resolved without impacting performance will require thoughtful database schema design.

Next Steps:

- Collaborate with the data engineering team to investigate the root causes of identified quality issues.
- Develop a clear standard for handling missing, invalid, and duplicate data.
- Align with you and other stakeholders on prioritization for these improvements.

Please let me know if you have additional insights to help guide our next steps.

Thanks,
Manoj