Hello Stakeholder,

I wanted to share key findings from the data quality review and outline next steps to ensure our data supports accurate and reliable insights. Below is a high-level summary of the analysis and recommendations:

Key Questions About the Data

1. **Mandatory Fields**:
   - Should fields like `state`, `role`, and `signUpSource` always be populated in the user and transaction datasets?
   - What are the minimum acceptable levels of missing or invalid data for production use?
2. **Future Dates and Invalid Timestamps**:
   - Are there specific business rules for handling future or invalid dates in fields like `createdDate` and `lastLogin`?
3. **Duplicate Records**:
   - Should duplicate records (e.g., same `user_id` or `barcode`) be resolved, flagged, or retained for further review?

How I Discovered Data Quality Issues

By conducting a detailed analysis using statistical methods and visualizations, I identified the following significant issues:

1. **Missing Data**:
   - Several key columns, such as `role` and `state`, have missing values.
   - Missing `barcode` values in the items table make it challenging to join with the brands table, resulting in incomplete information.
2. **Duplicate Entries**:
   - Duplicates in the user and brand datasets could lead to inflated counts and inaccurate reporting.
3. **Inconsistent Formats**:
   - Fields like `categoryCode` and `barcode` show inconsistencies, which complicate analysis and integrations.
   - Data types such as `pointsEarned` appearing as floats instead of integers can lead to misinterpretation.

What I Need to Resolve These Issues

1. **Business Context**:
   - Clarification on which fields are mandatory and acceptable levels of missing or invalid data.
2. **Data Source Knowledge**:
   - Understanding how data is ingested and stored (e.g., APIs, manual uploads) to pinpoint the origins of errors.
3. **Duplicate Handling**:
   - Guidance on whether duplicates should be removed, flagged, or treated as valid.

Additional Information Needed

1. **Data Lineage**:
   - A clearer view of the end-to-end flow from data source to analysis to identify bottlenecks.
2. **Expected Volume and Frequency**:
   - Information on data ingestion frequency and anticipated growth to inform scaling strategies.
3. **Usage Scenarios**:
   - Specific business questions or KPIs that the data should address to align the models with organizational objectives.

Anticipated Performance and Scaling Concerns

1. **Data Volume Growth**:
   - Increasing volumes may lead to slower queries. I propose implementing indexing and partitioning strategies in production to address this.
2. **Automated Data Cleaning**:
   - Adding validation pipelines to resolve issues (e.g., duplicates, invalid dates) could increase processing time. Prototyping will help balance accuracy and efficiency.
3. **Duplicate Management**:
   - Flagging or resolving duplicates without impacting performance requires thoughtful database schema design.

Next Steps:

- 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