Subject: Insights on Data Quality and Next Steps

Hi [Product Manager/Business Stakeholder],
I hope you're doing well! I am Pratik, and I am a Data Analyst on the Alpha team. My responsibilities as a Data Analyst include analyzing data, ensuring its quality and creating robust and scalable data models for the warehouse. The data analytics team is currently working towards our new data warehouse and reporting capabilities, and as part of this, I recently conducted a preliminary analysis of our raw datasets - Receipts, Users and Brand data.

Overview of Analysis

In this data quality check, my primary focus was to get an understanding of the completeness, validity, and consistency of our datasets. This process involved an exhaustive review of the data structure - identifying missing or inconsistent entries, and checking for any outliers that might affect our analyses and ultimately the new data models. The table below summarizes the key findings from the analysis and follow-up questions that will assist me and the team to address these issues effectively and optimally.

Key Findings

Quality Factor	Objective	Approach	Observation	Impact	Example	Question
Completeness	Identifying gaps in data that could affect analysis.	Calculate the % of missing values in the fields.	Significant missing values in several key fields	This impacts analysis of user purchasing behavior and trend analysis.	`purchaseDate` in Receipts dataset has 40.04% missing values.	Why is there a high percentage of missing values in critical fields?
Validity	Ensure that data points are valid and within expected domain ranges.	Check for invalid values and compare actual data types with expected types.	Invalid state codes in the Users dataset.	Affects geographic analysis and user activity.	`state` field in Users dataset has multiple invalid entries like 'nan'.	What is the process for validating state codes during data entry?
Consistency	Data is logically consistent across different attributes.	Review the relationships and dependencies amongst data points.	Inconsistent `rewardsReceiptStatus` values in Receipts dataset.	Challenges in accurately tracking receipt statuses.	`rewardsReceiptStatus` field shows several statuses that do not align with our defined statuses.	Are there any specific business rules affecting the `rewardsReceiptStatus` field that we need to be aware of?

Outliers	Identify data points that significantly deviate from the norm.	Use statistical methods to detect unusual values.	Significant outliers in pointsEarned, purchasedItemCount, and totalSpent fields.	Potential data entry errors or unique user behaviors affecting analysis.	`totalSpent` field in Receipts dataset shows both very high and very low outliers.	Are there known user behaviors or patterns that could explain these anomalies?

Clarification Needed

To understand these data issues, I have a few questions listed below with some examples. Any insights into this would really help us in the process:

- **Data Collection Process**
  Example: In the Receipts dataset, the purchaseDate attribute has 40%+ missing records. Understanding the data ingestion process could help us identify why there are so many missing values.
  Question: What is the process for data generation and ingestion? Are there any manual steps in this pipeline that could introduce errors or incomplete data?

- **Data Cleaning Procedures**
  Example: In the Users dataset, the State attribute has several records of ‘nan’ entries. We need to understand the data cleaning processes that could help us implement a robus logic to handle these errors.
  Question: What data cleaning and validation rules are currently employed before or after data is ingested into our system? How are invalid entries handled?

- **User Behavior**
  Example: The totalSpent field in the Receipts dataset has both high and low outliers, found through statistical tests. Have we performed any specific user behavior research such as bulk purchases or frequent small transactions, that will supplement an explanation?
  Question: Are there any known user behaviors or patterns that may explain the data anomalies we observed, such as that in the totalSpent field?

- **Business Rules**
  Example: The Receipts dataset has another attribute called rewardsReceiptStatus that shows several statuses that do not align with the set domain of the statuses. Understanding any special business rules around status assignment could clarify these inconsistencies.
  Question: Are there any specific business rules or logic that could affect data validity or consistency? How should these be reflected in the data?

- **Additional Details on Null Values**
  Example: For the pointsAwardedDate field in the Receipts dataset, which has about 52% missing values, we need to understand the reasons for these gaps to help in formulating imputation strategies.

Question: What are the underlying reasons for high null values in certain fields? Are there specific actions taken when data is missing?

Discovering Data Quality Issues

These issues were identified through thorough data quality checks, such as analyzing null values, performing validity and consistency checks, and detecting statistical outliers. These steps help ensure that our data remains accurate, consistent, and reliable for analysis and reporting.

Steps to Resolve Data Quality Issues

To resolve these issues, I need further insights into:

- The data collection, validation, and maintenance processes.
- Existing standard operating procedures for data entry and validation.

Additional Information Needed

To optimize our data assets, it would be helpful to gather some further details:

- **Anticipated Changes or Additions to Data Fields**
  Understanding any upcoming changes or planned additions to the data fields will help us design a flexible and scalable data model that can accommodate future needs.
- **Plans for Integrating this Data with Other Datasets or Systems**
  Knowing how this data will integrate with other datasets or systems ensures that we can create a cohesive and comprehensive data architecture, facilitating better data interoperability and analytics.
- **Feedback from Users Who Interact with These Data Points Regularly**
  Gathering insights from users who frequently interact with these data points will help us understand their challenges and needs, allowing us to make targeted improvements that enhance data usability and reliability.

Performance and Scaling Concerns

A secondary concern is the performance and scalability of the warehouse models as these errors could not only cause pipeline issues, but also make it inefficient for analysts to query and provide high quality insights.

- Inefficient queries and longer processing times due to missing and inconsistent data.
- Ensuring the data model can handle increasing volumes of data without performance degradation.
- Implementing robust validation rules to prevent erroneous data from affecting the production environment.

Next Steps

I would appreciate your feedback on the findings above and any additional insights you may have about the data issues that will help us. Additionally, I believe a collaborative session between your team and other stakeholders to discuss these issues will be a great step toward formulating a resolution plan. While we are in the process of building our warehouse, setting up temporary data monitoring pipelines will provide us with a more thorough understanding of data captures and its nature.

Thank you for your time and attention. Your insights and feedback will be invaluable in enhancing our data quality and ensuring reliable, actionable insights for our business.

Best regards,
Pratik Watwani
Data Analyst