Fourth: Communicate with Stakeholders

4. Write a short email or Slack message to the business stakeholder

Subject: Data Integrity Analysis – Key Issues & Business Impact

Hi [Name],

I recently conducted an analysis on our user activity, receipts, and purchase data to ensure accuracy and reliability in our reporting. During this process, I identified a few critical data integrity issues that could impact our ability to make accurate business decisions. Below are key findings and their impact:

- We found that high-value transactions are missing receipt records, and several receipts contain user_id values that do not exist in our user's table.
  Business Impact: This could potentially affect our ability to associate transactions with users accurately, which leads to an underreporting of revenue and customer segmentation and retention metrics could be inaccurate.
- Some records were duplicated, particularly in receipts, where the same receipt appeared multiple times under different user entries.
  Business Impact: This could result in inflated metrics, especially in reports for receipt volume and spending.
- There are thousands of unique barcode values that don't match known product brands.
  Business Impact: If barcode data isn't standardized, it could lead to product catalog mismatches and inefficient supply chain tracking, impacting operational efficiency.
- During the review, I have identified logical inconsistencies in the receipt status field. One key issue found is that some receipts marked as "Rejected" still have points awarded to users, which contradict our business rules.

Next Steps and Action Plan:
Address the identified issues by removing or correcting critical missing and duplicating data points.
Collaborate with the Engineering team to investigate data pipelines and ensure all transactions are accurately linked.
Implement automated anomaly detection to flag abnormal transactions proactively.

I would love to discuss this further to prioritize the most impactful fixes. Let me know when we can connect. Looking forward to hearing from you.

Best Regards,
Nikhitha Jagarlamudi

I hope the questions outlined in the document for Task 4 have been thoroughly addressed in the email above. Further elaboration on these questions can be provided during discussions when the business owner connects with the team.

- **What questions do you have about the data?**
  What is the expected relationship between user_id and receipts? Should every receipt have a valid user_id, or are there cases where standalone receipts exist?
  What is the source of barcode data? Are we missing an external reference list to validate product brands?
  Can we confirm that the receipt statuses (such as Accepted, Rejected, and others) are always set based on clear, consistent rules? Are there instances where the logic for assigning status has failed, such as a Rejected receipt having points awarded?

- **How did you discover data quality issues?**
  The data quality issues were identified through systematic approach combining data validation checks, exploratory data analysis (EDA), and anomaly detection.
  Missing Data Analysis: By running python scripts to check for NULL or missing values in critical fields found high-value receipts were disproportionately missing, also discovered incomplete records that could hinder our ability to link transactions with users and brands.
  Data Duplication: I ran script to check for duplicate entries, particularly in the users table, where the same data have been recorded multiple times
  Foreign Key Validation: Detected receipts with user_id values that don't exist in the user's table.
  Categorical Consistency Checks: Barcode values did not match known product brands, indicating potential data standardization issues.

- **What do you need to know to resolve the data quality issues?**
  Data Pipeline Flow: How and when is data ingested from different sources?
  Product Data Source: Is there a master list of barcodes/brands we should be validating against?
  Business Rules for Receipt Status: Can we have a clear set of rules for when a receipt should be marked?

- **What other information would you need to help you optimize the data assets you're trying to create?**
  Business Priorities: Which data quality issues have the highest impact on revenue, operations, or customer experience?
  User Behavior Insights: Are users failing to upload receipts, or is the system failing to capture them?

Historical Data Consistency: Are data quality issues consistent over time, or are they recent anomalies?

- **What performance and scaling concerns do you anticipate in production and how do you plan to address them?**

  Large Dataset Handling: As the volume of receipts, users, and brands increases, running complex SQL queries may slow down. To address this, we'll need to consider indexing key fields.

  Data Duplication and Integrity: With larger datasets, data duplication could become a bigger problem, leading to incorrect reporting and inconsistent analysis.

  Real-time Data Sync: If we move to a real-time data pipeline for receipt and user data, we'll need to handle high-throughput data flows efficiently.

  Anomaly Detection in Streaming Data: Need to detect flag issues as transactions are ingested.