Hi Product/Business Leader,

I’ve completed an initial analysis of the receipts, users, and brands datasets to ensure they’re reliable for business decisions. Below is a summary of findings, questions, and next steps to improve data quality.

**1. Key Data Quality Issues & Questions**

**Users Data**

* **Issue:** 283 duplicate user IDs out of 495 records. (e.g., 20 entries for the same user).
* **Question:** Are these duplicates intentional (e.g., test accounts) or errors? Should we merge or delete them? Do we create a new record everytime we update information of a user?
* **Missing Information:** Key fields like lastLogin (12.5% missing), signUpSource (9.7% missing), and state (11.3% missing) have gaps.
* **Question:** How important are these fields for our analysis or customer-facing features?

**Brands Data**

* **Issue:** examples like 20% brands missing brandCode, 13.28% missing category, and 52.4% missing topBrand flag.
* **Question:** Are these critical for our product recommendations or marketing efforts?
* **Issue:** Duplicate barcodes (e.g., barcode 511111004790 linked to both “Alexa” and “Bitten Dressing”).
* **Question:** Are barcodes reused intentionally (e.g., for parent/child products), or is this a data entry error?

**Receipts Data**

* **Issue:** 40% of receipts lack purchaseDate. Should we infer it from createDate?
* **Issue:** In some cases, the purchaseDate is after the createDate. Is this an error, or does it reflect a specific process
* **Issue:** 38.9% of receipts have missing totalSpent, limiting purchase insights.
* **Question:** Why does review status ‘SUBMITED’ responsible for nearly all missing entries? What might cause these to be missing? What is the difference between ‘FINISHED’ and ‘SUBMITTED’ in the rewardReceiptStatus?

**Rewards Items Data**

* **Issue:** 55% of items lack barcodes, making it hard to link purchases to brands.
* **Issue:** Price discrepancies (e.g., total (final prices \* quantityPurchased) does not match the totalSpent in some receipts).
* **Question:** How should we calculate total spend per receipt? What is the relationship between itemPrice, finalPrice and discountedItemPrice?

**2. How Issues Were Discovered**

To assess the data, I loaded the JSON files into Python, which helps analyze datasets efficiently. Here’s what I did:

* **Checked for Missing Values**: Calculated how many entries were blank in each column and what percentage that represents.
* **Looked for Duplicates and inconsistencies in unique IDs**: Identified repeated records, like user IDs or barcodes, to see if they align with expectations.
* **Data mismatches**: Calculated the differences between expected totals and computed sums per receipt.
* **Compared Dates**: Examined purchaseDate and createDate in receipts to spot any unusual patterns.

These steps revealed the gaps and inconsistencies I mentioned above, which could affect how reliable our data is for decision-making or customer features.

**3. Additional Information Needed**

* **Data Dictionary:** Definitions for ambiguous fiel (e.g., finalPrice, discountedFinalPrice, partnerItemId).
* **Data Documentation:** A guide explaining what each field means, how it’s collected, and any known quirks or limitations.
* **Source System Details:** How is data collected (e.g., app workflows, third-party integrations)?
* **Data Ownership:** Information on which teams or sources are responsible for these data sets to streamline corrective actions.
* **Business Rules:** Insight into the logic behind fields like bonusPointsEarnedReason or rewardsReceiptStatus. For example, why do some receipts earn points and others don’t?

**4. Performance & Scaling Concerns**

As we prepare to use this data in a live environment, I foresee a few challenges:

* **Growing Data Volume**: As more users submit receipts, our dataset could expand rapidly, slowing down our systems if we’re not prepared.
* **Cleaning Overhead:** With the current level of missing or inconsistent data, preparing it for use could take significant time and effort.
* **Real-Time Needs:** If we want features like instant rewards, we’ll need to process receipts quickly, which could strain our current setup.

To tackle these, I plan to:

* **Automate Data Checks:** Build tools to catch and fix issues (e.g., missing values or duplicates) as data comes in, reducing manual work.
* **Optimize Performance:** Use efficient database techniques (like indexing) to speed up data retrieval, even as the dataset grows.
* **Scale with Technology:** Leverage tools like Apache Spark to process large volumes of data quickly, ensuring we can handle growth without hiccups.

**Next Steps:**

1. Schedule a 30-minute discussion to clarify business rules for duplicates and missing data.
2. Share a data dictionary to align on field definitions.
3. Pilot fixes for high-impact issues (e.g., user deduplication) and validate results.

Let me know a time that works for you!

Fan

Best