Hello,

I have completed an initial review of the provided datasets and identified several data quality issues that could impact analysis. Below is a summary of key findings, questions, and proposed next steps.

Summary of Data Quality Findings:

1. Improper JSON Structure:
   - The JSON files (brands.json, receipts.json, and users.json) are not structured as single valid JSON objects. They contain multiple JSON objects without proper array encapsulation, requiring line-by-line parsing. This increases complexity and the risk of data integrity issues.
2. Missing Data:
   - Significant missing fields across datasets, including ~4,600 missing fields in Receipts and ~1,651 in Brands.
   - Key fields like totalSpent, pointsEarned, and rewardsReceiptItemList had notable gaps.
3. Duplicate Records:
   - Identified 283 duplicate user records that may affect user-based analytics.
4. Invalid References:
   - 148 receipts reference user IDs not present in the Users dataset.
   - Over 7,000 receipt items have barcodes not matching any entries in the Brands dataset.
5. Data Type and Format Issues:
   - Fields such as totalSpent and pointsEarned are stored as strings in some cases, affecting numerical analysis.
   - Detected outliers in fields like totalSpent and quantityPurchased.
6. Inconsistent Categorical Data:
   - Missing or inconsistent values in the Users' state field.
   - Variability in rewardsReceiptStatus values, requiring clarification.

Questions Regarding the Data:

1. Are all rewardsReceiptStatus values valid, or should certain statuses be excluded from analysis?
2. Should receipts missing rewardsReceiptItemList be considered valid, or do they represent incomplete data?
3. What is the preferred method for handling duplicate user records—merging based on common fields or removal?
4. Are there valid cases where barcodes in receipts would not exist in the Brands dataset?
5. Is there a plan to standardize JSON formatting for future data exports to ensure proper structure?

Information Needed to Resolve Issues:

1. Business rules for handling missing and duplicate data.
2. Clarification on acceptable thresholds for outliers (e.g., high quantityPurchased values).
3. Any existing mapping files that accurately link barcodes to brands.
4. Guidance on whether the current JSON structure should be corrected at the source.

Performance and Scaling Considerations:

1. Data Volume: Indexing key fields like userId, receiptId, and barcode would improve query performance as data scales.
2. Scaling Strategy: Considering a distributed data warehouse (e.g., Redshift, BigQuery) could help manage growing data volumes.
3. Data Validation: Implementing automated validation checks to flag missing data, invalid references, and outliers before data enters production.

Next Steps:

Please let me know if you'd like to discuss these findings further. Once I have clarification on the questions above, I can proceed with data cleaning and optimization.

Thank you for your time, and I look forward to your guidance.

Best regards,