1. Receipt Table
Issues:
1. Dupes in `user_id`:
   - Identify duplicate `user_id` entries and determine if they are valid (e.g., same user scanning multiple receipts).
   - If invalid, remove duplicates or merge records if necessary.

2. Invalid `receipt_id`:
   - Validate `receipt_id` format (e.g., length, characters, uniqueness).
   - Remove or flag invalid entries.

3. Invalid `bonusPointsEarned`:
   - Ensure `bonusPointsEarned` is a non-negative number.
   - Replace invalid values with defaults (e.g., 0) or remove the record.

4. Invalid or null `createdDate`/`dateScanned`:
   - Validate date formats and ensure they fall within a reasonable range (e.g., not future dates).
   - For null values, impute based on other data (e.g., use `dateScanned` if `createdDate` is null) or remove the record.

5. Invalid `rewardStatus` values:
   - Define valid values for `rewardStatus` (e.g., "Finished", "Pending").
   - Replace invalid values with a default (e.g., "Unknown") or remove the record.

6. Invalid `totalSpent`:
   - Ensure `totalSpent` is a non-negative number.
   - Replace invalid values with defaults (e.g., 0) or remove the record.

---

2. Brand Table
Issues:
1. Dupes in `barcode` but different `category`:
   - Investigate if the same barcode is associated with multiple categories.
   - Resolve conflicts by updating the category or flagging the record for review.

2. Missing spellings in `CPG` column (e.g., "Cogs"/"Cpgs"):
   - Standardize the `CPG` column values (e.g., "CPG").
   - Use string matching or regex to correct misspellings.

3. Null in `category` or `categoryCode`:
   - Impute missing values based on other data (e.g., use the most common category for the brand).
   - If no data is available, flag the record for review.

4. Invalid or null `names` (e.g., "test. brand"):
   - Remove or flag records with invalid or placeholder names.

- Standardize brand names (e.g., trim whitespace, capitalize correctly).

5. Null in `topBrand` column:
- Impute missing values based on brand popularity or other metrics.
- If no data is available, set a default value (e.g., "False").

6. Invalid or null `brandCode` (e.g., "TEST BRANDCODE @1612366146051")
- Validate `brandCode` format (e.g., length, characters).
- Replace invalid or placeholder values with a standardized code or remove the record.

---

3. User Table
Issues:
1. Dupes in `user_id`:
- Identify duplicate `user_id` entries.
- Merge records if they represent the same user or remove duplicates if invalid.

---

4. Reward Item Details (Separated from `rewardsReceiptItemList`)
Issues:
1. Nulls in multiple columns (`needsFetchReview`, `PartnerItemId`):
- Impute missing values based on other data (e.g., set `needsFetchReview` to "False" if no data is available).
- Flag records with critical missing data for review.

2. Missing `barcode` (e.g., `originalReceiptItemText: "PEPSI COLA"`):
- Use `originalReceiptItemText` to infer or lookup the `barcode`.
- If no `barcode` can be determined, flag the record for review.

Key Data Relationships and Implications:

One-to-Many Relationships: As you know, there are one-to-many relationships between the tables:
- A single receipt can have multiple items (represented in `RewardItemDetails`), so ensuring that `receipt_id` and `barcode` are correct and unique is crucial.
- A user can scan many receipts over time, which means we need to ensure accurate linking of users to receipts.
- A barcode may appear across multiple items in different receipts, so ensuring the `barcode` in the `Brand` table is unique and correctly categorized is essential for accurate brand performance tracking.

Next Steps:

1. Clean Up the Invalid Data:

- Validate and standardize `user_id` and `receipt_id` formats.
- Fix invalid `totalSpent`, `bonusPointsEarned`, and `rewardReceiptStatus` values.
- Address missing and invalid dates (`createDate`, `dateScanned`) to ensure proper event tracking.

2. Fix Duplicate and Inconsistent Data:
   - Resolve duplicates in the `barcode` column in the `Brand` table and ensure consistent category assignments.
   - Clean up null or invalid `brandName`, `category`, and `categoryCode` fields in the `Brand` table.
   - De-duplicate `user_id` entries in the **User** table.
3. **Data Normalization and Enrichment:**
   - **Ensure that all `barcode` values are present in the `RewardItemDetails` table and are properly linked to the correct brand.**
   - **Fill missing values in critical fields like `needsFetchReview` and `partnerItemId` in `RewardItemDetails`.**

Performance and Scaling Concerns:

- Volume of Data: As we scale, the volume of records will increase. Ensuring that our data validation and cleaning processes are efficient will be critical to avoid bottlenecks, especially during monthly reporting cycles.

- Query Performance: Queries that join large tables like `Receipts` and `RewardItemDetails` may become slower as the data grows. We'll need to consider optimizing these queries with proper indexing or denormalization strategies to ensure consistent performance.

- Data Consistency in Production: Implementing a more rigorous data validation pipeline (e.g., via ETL processes) will help catch these issues early in production and prevent them from affecting live reporting.

Emails to Business:

Data Quality Issues Identified – Action Needed

Hi - Name of Business User,

I hope this message finds you well. As part of our ongoing efforts to ensure the accuracy and reliability of our data, I've conducted a thorough review of the datasets we're working with. I've identified several data quality issues that need to be addressed to ensure our analyses and business decisions are based on clean, consistent, and trustworthy data.

Here's a high-level summary of the issues:

1. Receipt Table
- Duplicate User IDs: Some users appear multiple times, which could skew user-level metrics.
- Invalid Receipt IDs: Some receipt IDs don't match the expected format or are missing.
- Invalid Bonus Points Earned: Some entries have negative or nonsensical values.
- Invalid or Missing Dates: Some `createdDate` or `dateScanned` values are either invalid or null.
- Invalid Reward Status: Some entries have unexpected values in the `rewardStatus` field.
- Invalid Total Spent: Some values are negative or unrealistic.

2. Brand Table
- Duplicate Barcodes with Different Categories: The same barcode is associated with multiple categories, causing inconsistencies.
- Misspelled CPG Column: Variations like "Cogs" or "Cpgs" need standardization.
- Missing or Invalid Category Data: Some entries have null or invalid `category` or `categoryCode` values.
- Invalid or Null Brand Names: Placeholder names like "test. brand" need to be cleaned.
- Missing Top Brand Flags: Some entries have null values in the `topBrand` column.
- Invalid Brand Codes: Some codes are placeholders (e.g., "TEST BRANDCODE @1612366146051").

3. User Table
- Duplicate User IDs: Some users appear multiple times, which could lead to inaccurate user counts.

4. Reward Item Details
- Missing Barcodes: Some items don't have barcodes, relying only on text descriptions (e.g., "PEPSI COLA").
- Null Values in Key Columns:** Columns like `needsFetchReview` and `PartnerItemId` have missing data.

Questions for the Team
To resolve these issues effectively, I need your input on the following:
1. Are there specific business rules or thresholds for validating fields like `bonusPointsEarned`, `totalSpent`, or `rewardStatus`?
2. How should we handle duplicate barcodes with conflicting categories? Should we prioritize one category or flag them for manual review?
3. For missing or invalid brand names, should we remove these records or attempt to infer the correct names?
4. Are there any known issues or edge cases in the data collection process that could explain some of these anomalies?

Next Steps
1. Data Cleaning: I'll start cleaning the data based on the above issues, using standardized rules and imputing missing values where possible.

2. Validation Rules: I'll define and document validation rules to prevent similar issues in the future.
3. Automated Checks: I'll implement automated data quality checks to flag issues as new data comes in.

Other Questions:
As we move toward production, we'll need to address:
- Data Volume:With more users and receipts, we'll need efficient indexing and partitioning strategies to handle large datasets. Is this monthly file or weekly file or daily file?
- Validation: Can someone in the business team help to validate the rules file?