Explore the data

Review the unstructured csv files and answer the following questions with code that supports your conclusions:

Data Cleaning, Exploration and Visualization is performed in Python in Python before importing data into SQL

Are there any data quality issues present?

- **Missing/Duplicate Values**: The data contains missing values across multiple fields. Removed duplicate entries based on the BARCODE column to maintain primary key integrity and fixed any null BARCODE value and filled missing BIRTH_DATE values with a default older date (1900-01-01 00:00:00) and replaced missing BARCODE (Foreign key)values in Transactions with 0 to maintain data integrity.
- **Inconsistent Date Formats**: Columns such as CREATED_DATE, BIRTH_DATE, PURCHASE_DATE, and SCAN_DATE have inconsistent date formats.
  Converted all date-time values into a SQL-friendly format (YYYY-MM-DD HH:MM:SS) for consistency.
- **Numeric Issues**: Columns like quantity and sales contain problematic values such as "NaN," "zero," and scientific notation.
  Replaced zero values with missing values, fixed the numeric issues, and aligned the price and quantity values correctly across rows to prevent duplicates.
- **Encoding Issues**: Non-ASCII characters in columns like CATEGORY_2, CATEGORY_3, CATEGORY_4, MANUFACTURER, BRAND in Products, and STORE_NAME in Transactions cause encoding issues during SQL uploads.Used unicodedata.normalize() to replace non-ASCII characters and resolve encoding issues.

Are there any fields that are challenging to understand?

- **Category Hierarchy**: It appears that product categories follow a structured hierarchy from category_1 to category_4, but the exact relationship between these levels needs to be confirmed.
- **Quantity**: For products with fractional quantities, I used they are measured by weight, a quantity of 1 is assigned based on a specific weight threshold.
- **Sales**: The "Sales" field represents rewards in USD from the Fetch app, rather than actual transaction amounts as amounts ranging from a $0.01 to $462.82 (more concentrated towards lower value)

Data Visualization Insights (Python)

- **Language Distribution**: English is the dominant language, with Spanish (es-419) users suggesting a potential for bilingual marketing.
- **Gender Distribution**: Approximately 70% of users are female, presenting opportunities for targeted marketing, particularly in lifestyle, fashion, and health sectors.
- **Age Distribution**: Users aged 24-26 make up the largest group, indicating they could influence future product features, promotions, and social media strategies.
- **Top Store Analysis**: Walmart leads in transaction volume, highlighting its importance for strategic partnerships and marketing efforts.
- **Category Analysis**: Health & Wellness and Snacks dominate, reflecting growing demand in these areas. Future offerings could cater to this interest.
- **Transaction Trends**: Transactions from June 12 to September 8, 2024, suggest seasonality or promotional events, though data availability issues may impact accuracy. This insight can guide future marketing campaigns.
- **Days Between Purchase and Scan Dates**: The average 2-day gap indicates users are quick to upload receipts, suggesting the potential for timely reminders to enhance engagement.
- **State Distribution**: Texas, Florida, and California have the highest number of users, providing opportunities for region-specific marketing and partnerships.

Assumptions

- **Primary Key Integrity**: Assumed removing duplicates and NULL based on primary keys suffices for data cleaning, and non-primary key columns can be null (storage or collection issue).
- **Foreign Key Integrity**: Missing BARCODE in Transactions is replaced with 0 to ensure no null values for foreign keys.
- **Handling Missing Data**: Missing BIRTH_DATE in Users filled with a default old date (1900-01-01 00:00:00) to maintain integrity.
- **Date Formatting**: Standardized all date columns to SQL-friendly format (YYYY-MM-DD HH:MM:SS).
- **Non-ASCII Characters**: Non-ASCII characters in product and transaction tables were replaced using unicodedata normalization to avoid encoding issues.
- **Category Hierarchy**: Assumed categories follow a hierarchy (category_1 to category_4), but structure requires further clarification.
- **Purchase vs. Scan Dates**: Assumed purchase_date is the actual transaction date, and scan_date is when it was reported in the Fetch app.
- **State Field**: Null values in the state field are not critical unless they stem from missing user input.
- **Quantity**: Products measured by weight are assigned a quantity of 1 for a specified weight threshold, explaining the presence of fractional quantities.
- **Sales**: “Sales” refers to rewards in USD from the Fetch app, distinct from actual transaction amounts.
- **Language**: The system supports English (“en”) and Spanish (“es-419”), which may impact user segmentation and content localization.