**Fetch User Sales and Product Analysis**

Presented by: **Ram Charan Shashank Jangam**

**Overview:**

The dataset provided consists of three CSV files: products, transactions, and users, each detailing different aspects of a retail environment. The **products** file contains information about various product categories, manufacturers, brands, and barcodes, giving insights into the product taxonomy and branding. The **transactions** file records individual purchase events, including receipt IDs, purchase and scan dates, store names, user IDs, and details about products and quantities purchased. The **users** file provides demographic information about the customers, such as their account creation dates, birthdates, states of residence, preferred languages, and gender.

**Understanding Entity Relationship Model:**

A screenshot of a diagram

Description automatically generated

The entity-relationship (ER) model depicts the structure and relationships among the datasets provided. It comprises three main entities: **Users**, **Transactions**, and **Products**. The **Users** entity includes demographic and account information such as user IDs, creation dates, birth dates, states, languages, and gender. The **Transactions** entity captures details of each purchase, linking it to users and products through fields like receipt ID, purchase and scan dates, store names, and barcodes. The **Products** entity contains product classifications, manufacturers, brands, and barcodes.

This ER model is essential for understanding how these entities interact, allowing us to join data tables based on shared keys (e.g., user IDs and barcodes). It guides the structuring of our code and ensures that analyses accurately reflect user behavior, product sales, and transaction details, thus helping to extract meaningful insights from the dataset.

**PART 1: DATA EXPLORATION   
1) Are there any data quality issues present?**

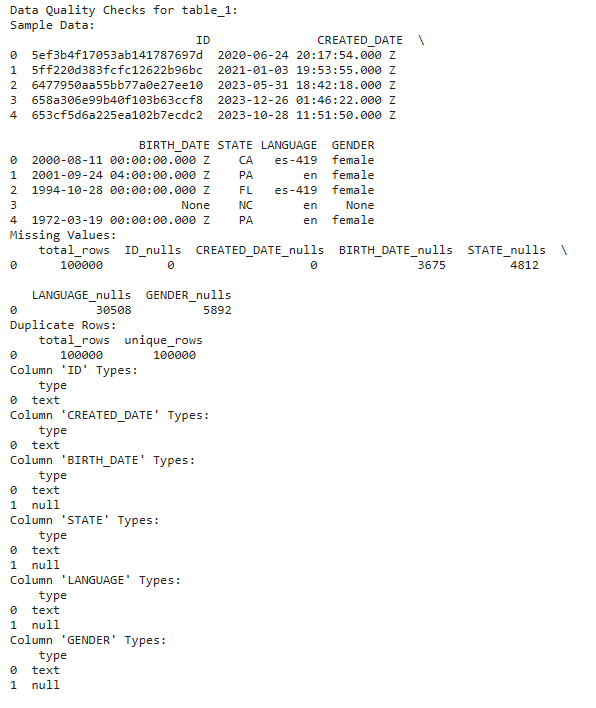
The analysis of the three datasets has revealed various data quality issues across all tables. Below is a summary of the identified problems for each dataset, accompanied by relevant codes and screenshots:

**Data quality checks for all tables**  
A screenshot of a computer code

Description automatically generated

**TABLE 1: USER DATA**

* **Missing Values**: Out of 100,000 rows, there are significant missing values in critical fields:
  + **BIRTH\_DATE**: 3,675 nulls
  + **STATE**: 4,812 nulls
  + **LANGUAGE**: 30,508 nulls
  + **GENDER**: 5,892 nulls
* **Inconsistent Data Types**:
  + Key fields like **CREATED\_DATE** and **BIRTH\_DATE** are stored as text instead of the appropriate datetime format. This misalignment complicates date-related calculations and may lead to inaccurate analyses.
* **Inconsistencies in Entries**:  
  The "Gender" column exhibits inconsistencies in how responses are recorded. Some entries use the phrase "Prefer not to say," while others are recorded as "prefer\_not\_to\_say." These variations can lead to challenges in data analysis and aggregation, as they may be interpreted as distinct categories despite conveying the same meaning.

**OUTPUT:**  


**INCONSISTENT VALUES USED IN THE GENDER COLUMN:**A screenshot of a computer

Description automatically generated

**TABLE 2: PRODUCT DATA**

* **Missing Values**: There are numerous null entries across key fields, including:
* **CATEGORY\_1**: 111 nulls
* **CATEGORY\_2**: 1,424 nulls
* **CATEGORY\_3**: 60,566 nulls
* **CATEGORY\_4**: 778,093 nulls
* **MANUFACTURER**: 226,474 nulls
* **BRAND**: 226,472 nulls
* **BARCODE**: 4,025 nulls
* **Data Type Issues**: The **BARCODE** field is classified as "real," despite containing both floating-point and integer representations (e.g., 7.28048E+11 and 22600601012). It should be standardized as an integer to avoid processing complications. Additionally, the **BARCODE** field, which is intended to be a unique identifier (primary key), contains duplicate entries. This violation of uniqueness further compromises the integrity of the dataset and can lead to inaccuracies in data retrieval and analysis.

**OUTPUT:**A screenshot of a computer

Description automatically generated

**TABLE 3: TRANSACTION DATA**

**OUTPUT:**

A screenshot of a computer

Description automatically generated

* **Missing Values:** Among the 50,000 rows, the dataset has one critical area of concern:

**BARCODE:** 5,762 null entries, which can affect product identification and transaction integrity. All other fields are complete.

* **Data Type Issues: PURCHASE\_DATE** and **SCAN\_DATE** are stored as text but should be formatted as datetime for accurate date operations.

**BARCODE** is categorized as "real," but it should be stored as an integer to properly reflect its unique identifier role.

**FINAL\_QUANTITY** and **FINAL\_SALE** are recorded as text, whereas they should be numeric to facilitate calculations and analyses.

* **Inconsistencies in Entries:**  
  There are inconsistencies in the entries for the **STORE\_NAME** and **FINAL\_QUANTITY** fields. Standardizing these entries is essential to ensure data accuracy and facilitate reliable analysis.

**VISUALIZATIONS ( Distributions and Top Performers)**

**A graph showing the distribution of users

Description automatically generated**

1. Distribution of user's age

**A chart with numbers and text

Description automatically generated with medium confidence**

2. Distribution of products

1. **Are there any fields that are challenging to understand?**

**Understanding Challenges in Dataset Fields**

**TABLE 1: USER DATA**

* **STATE and LANGUAGE:** The STATE and LANGUAGE fields in the Users dataset use abbreviations and codes that are difficult to interpret without a reference key. To make effective use of these fields in analysis, additional context or decoding is needed.

**TABLE 2: PRODUCT DATA**

* **CATEGORY:**

It is difficult to understand the role of categories in forming a hierarchical tree for each product record, as there are no clear guidelines or captions provided. The CATEGORY\_1 to CATEGORY\_4 columns appear to represent different classification levels, but without a clear hierarchy, it's difficult to see how these categories connect to individual products.

* **BARCODE:**

The Bar Code field raises questions about uniqueness, as multiple products share the same barcode even though it is designated as the primary key in the dataset. This inconsistency poses issues regarding data integrity.

**TABLE 3: TRANSACTION DATA**

* **FINAL\_QUANTITY:**

There appears to be an inconsistency between the **FINAL\_QUANTITY** and **FINAL\_SALE** fields. Even when **FINAL\_QUANTITY** is recorded as zero, there are still values present in **FINAL\_SALE**. This discrepancy makes it challenging to interpret the relationship between these two fields, as it’s unclear if a quantity of zero should correspond to any sale amount. Additional clarification on how these fields interact would help improve data accuracy and understanding.  
Also, the FINAL\_QUANTITY field is difficult to interpret due to the presence of decimal values. This can be confusing and makes it harder to understand the quantity sold.

During my analysis of the **Products** and **Transactions** tables, I noticed missing values in the **Barcode** fields. Unlike other numeric fields, where it might make sense to fill in missing values with an average or mode, the Barcode field required a different approach for a couple of reasons:

* **Uniqueness**: Each barcode is a unique identifier tied to a specific product, so filling in missing values could create duplicates or incorrect associations.
* **Non-Integer Format**: The barcodes are stored in a non-integer format (like scientific notation or long numbers), making it impractical to replace missing values without impacting accuracy.

**Decision**: Given these factors, I chose not to fill missing values in the Barcode field. By leaving these as null, I maintained the uniqueness of each product identifier and avoided making unsupported assumptions about the data.

In the final CSV files, you’ll see that missing barcodes remain unfilled, reflecting the original data structure while preserving data integrity.

**PART 2: SQL QUERIES**

**Closed-ended questions:**

**1. What are the top 5 brands by receipts scanned among users 21 and over?**

This query identifies the five brands with the highest number of receipts scanned by users aged 21 and older. The calculation of user age is based on the difference between the scan date and the birth date. According to the results, **Equate** leads significantly, followed by **CVS**, **Pepsi**, and **Nature’s Bounty**. This suggests that these brands have strong engagement or visibility among adult users on the platform, potentially indicating popular choices within this demographic.

A screenshot of a computer

Description automatically generated

2. **What are the top 5 brands by sales among users that have had their account for at least six months?**

A screenshot of a computer

Description automatically generated

This query ranks the top five brands based on total sales among users who have held an account for six months or longer. The results show **Equate** as the leading brand in terms of sales, followed by **Pepsi**, **CVS**, and **Spring Valley Vitamins**. These findings suggest that long-term users prefer these brands, indicating potentially strong brand loyalty or frequent purchasing habits for these products over time.

**Open-ended questions:** **Which is the leading brand in the Dips & Salsa category?**A screenshot of a computer code

Description automatically generated

This query identifies the top-selling brand in the Dips & Salsa category by calculating total sales. The results show **Marketside** as the leading brand, with a total sales value of 196,900.06. This indicates that Marketside products are highly popular among customers in this category, making it a strong performer in the Dips & Salsa market. This insight could guide marketing strategies or product placement to maximize visibility and sales in this category.

**Assumptions**:

In this analysis, we interpreted transactions.FINAL\_SALE as a dollar amount reflecting a portion of the total transaction, where a float value (e.g., 3.5) represents a higher dollar figure, such as $3,500. For the open-ended question about identifying the leading brand in the Dips & Salsa category, we assumed that summing these final sale amounts would accurately indicate the brand’s popularity. We also assumed that the highest total sum of final sales would effectively identify the leading brand, reflecting consumer preference in this category.

**Part 3: Reporting to Stakeholders through email**

**Subject**: Data Quality Concerns and Request for Clarification

Dear Stakeholder,

Our recent data review highlighted several issues across the User, Product, and Transaction datasets, I’ll highlight the key data quality issues, interesting trends, and Requests for action in this email.

**Key data quality issues**

**1. Missing Values:** Many essential fields, such as dates, product categories, and barcodes, have missing data, affecting analysis completeness.

**2. Inconsistent Data Types:** Some fields are stored in unsuitable formats (e.g., dates stored as text), complicating calculations.

**3. Inconsistent Entries:** Similar entries, like gender options, appear in different formats, impacting data aggregation.

**4. Clarification on certain fields and Database normalization:** Additionally, we need clarification on certain fields, such as product category hierarchy, state/language codes, and the relationship between quantity and sales in transactions.

**5. Database Normalization issue:**During the normalization of the database tables, we identified several integrity constraint violations that prevented data from being loaded into the new tables following the defined foreign constraints. Specifically, certain products were missing barcodes, which led to transactions being recorded without associated barcodes. This resulted in null values that violate the foreign key constraints outlined in the ER diagram.

A screenshot of a computer program

Description automatically generated

**Interesting trends in the data**

This data analysis reveals valuable insights into our business performance. Understanding the dynamics of the top retail partners and user demographics is essential for informed decision-making and strategic growth

**1. Top Stores by Sales:** Walmart leads significantly, contributing over 50% of the total sales among the top 10 stores. This highlights Walmart as our most influential retail partner by a large margin.

Other stores like Costco, Sam’s Club, and CVS each hold smaller but notable shares, ranging from 5-9%. This distribution suggests that while Walmart is dominant, diversifying and strengthening partnerships with these other stores could provide additional growth.  
A pie chart with numbers and text

Description automatically generated

**3**. Top 10 stores by final sale

**2. User Demographics by State:** Texas, Florida, and California have the highest user counts, each with over 8,000 users. This concentration indicates that these states are key markets for us, where further marketing and customer engagement could yield significant returns.

States like New York and Illinois show lower but still substantial user counts, indicating additional areas where user engagement can be further optimized.  
  
**A graph of the states by user count

Description automatically generated**

**4**. Top 10 states by user count

**Request for action**

Additionally, we need clarification on certain fields such as product category, hierarchy, state/language codes, and the relationship between quantity and sales in transactions.

Could you provide any documentation or insights to help us address these issues? This information will enhance the accuracy of our analysis.

Thanks and regards,

Ram Charan Jangam