I. Users Dataset Documentation

1. Dataset Overview

The **Users Dataset** consists of **100,000 records** and provides demographic details about registered users, including their **birthdate, language preference, gender, and account creation date**.

Columns:

Column Name	Description
ID	Unique identifier for each user.
CREATED_DATE	Date the user account was created.
BIRTH_DATE	User's date of birth.
STATE	State where the user resides.
LANGUAGE	Primary language of the user.
GENDER	User's gender.

2. Data Quality Issues Identified

The following **data quality issues** were found during dataset exploration:

Missing Values

- **BIRTH_DATE**: **3,675 missing values**
- **STATE**: **4,812 missing values**
- **LANGUAGE**: **30,508 missing values**
- **GENDER**: **5,892 missing values**

These missing values may impact analysis, particularly for **demographic-based insights**.

Data Type Inconsistencies

- **CREATED_DATE and BIRTH_DATE were stored as objects** instead of `datetime`, requiring conversion for accurate date-based calculations.

Potential Data Anomalies

- The **GENDER** column contains **11 unique values**, including variations and ambiguous entries. Some values represent the same category but in different formats, while others indicate missing or unspecified gender preferences.

Identified Unique Values:

`"Female"`, `"male"`, `"non_binary"` and `"non-binary"` (same category but different formatting), `"transgender"`, `"prefer_not_to_say"` and `"prefer not to say"` (formatting inconsistency), `"not_listed"`, `"unknown"`, `"not_specified"`, `"my gender isn't listed"` (potential free-text entry)

  - **Formatting inconsistencies**: `"non_binary"` vs. `"non-binary"`, `"prefer_not_to_say"` vs. `"prefer not to say"`
  - **Free-text entries**: `"my gender isn't listed"`, `"not_listed"`
  - **Ambiguous values**: `"unknown"`, `"not_specified"`

- **STATE has 52 unique values**, aligning with U.S. states and territories but requiring validation.
- **LANGUAGE has only 2 unique values** (`"en"` and `"es-419"`, assumed to represent English and Spanish).

3. Data Cleaning & Standardization

To address these issues, the following cleaning steps were applied:

Handling Missing Values

- **BIRTH_DATE**: Left as `NaN` since imputation was not possible.
- **STATE**: Kept missing rather than assuming a default state.
- **LANGUAGE**: No assumptions were made, missing values retained as `NaN`.
- **GENDER**: No imputation was applied, and all unique values were retained.

Data Type Conversions

- **CREATED_DATE and BIRTH_DATE** converted to **`datetime`** format.
- **BIRTH_DATE** was used to calculate user age where necessary.

Standardizing String Values

- **STATE, LANGUAGE,** and **GENDER** standardized (capitalization, space removal).
- **GENDER** inconsistencies were analyzed but not forcefully corrected.

4. Assumptions & Considerations

During data cleaning and analysis, the following **assumptions** were made:

1. **Missing values were retained as NaN**, rather than applying assumptions for missing demographics.
2. **BIRTH_DATE was assumed to be correctly formatted** where present.
3. **CREATED_DATE was used as-is**, assuming it accurately reflects the user's registration date.
4. **STATE names were assumed to be U.S. states**, but further validation may be needed.

5. Conclusion

- The **Users Dataset** contains valuable demographic data but has missing values that may impact certain analyses.
- **Data type inconsistencies were corrected for dates**, and **string fields were standardized**.
- The dataset is now **cleaned and ready for analysis**, with missing values **retained where necessary to avoid assumptions**.

I. Transaction Dataset:

1. Dataset Overview

The **Transactions Dataset** consists of **50,000 records** and provides information on purchases made by users. The dataset includes key transaction details such as **receipt ID, purchase date, store name, user ID, barcode, final quantity, and final sale amount.**

Columns:

Column Name	Description
RECEIPT_ID	Unique identifier for each scanned receipt.
PURCHASE_DATE	Date when the transaction occurred.
SCAN_DATE	Date when the receipt was scanned into the system.
STORE_NAME	Name of the store where the purchase was made.
USER_ID	Identifier linking the transaction to a specific user.
BARCODE	Product barcode scanned from the receipt.
FINAL_QUANTITY	Total quantity of the purchased item.
FINAL_SALE	Total sale amount of the item.

2. Data Quality Issues Identified

The following **data quality issues** were found during the dataset exploration:

Missing Values

- **BARCODE**: 5,762 missing values (**~11.5% of transactions**).
- **FINAL_SALE**: 12,500 missing values (**~25% of transactions**).

Duplicate Records

- **171 duplicate transactions** were found and removed to ensure data consistency.

Data Type Issues

- **FINAL_QUANTITY was stored as a string** instead of a numeric value.
- **PURCHASE_DATE and SCAN_DATE were stored as objects** instead of `datetime`.

Outlier in `FINAL_QUANTITY` and `FINAL_SALE`

- A transaction had **`FINAL_QUANTITY = 276`** but **`FINAL_SALE = 5.89`** (missing in another instance), suggesting a possible outlier.
- The duplicate `RECEIPT_ID`, `STORE_NAME`, `USER_ID`, and `BARCODE` indicate a potential data capture issue.
- Assumed that **`FINAL_QUANTITY`** **represents weight/volume (e.g., grams or milliliters)** rather than item count.
- Data was **retained as valid**, considering unit-based pricing may explain the discrepancy.

3. Data Cleaning & Standardization

To address these issues, the following cleaning steps were applied:

Handling Missing Values

- **Missing BARCODE values**: Kept these transactions, which might to indicate potential scanning issues.
- **Missing FINAL_SALE values**: Left as NaN, since they cannot be inferred accurately.

Data Type Conversions

- **FINAL_QUANTITY converted to float** for accurate numerical operations.
- **PURCHASE_DATE and SCAN_DATE converted to datetime** for date-based analysis.

Duplicate Removal

- **171 duplicate records** were dropped to maintain transaction uniqueness.

4. Assumptions & Considerations

During data cleaning and analysis, the following **assumptions** were made:

1. **FINAL_QUANTITY as Volume or Weight-Based Transactions**
    - Since `FINAL_QUANTITY` contains **decimal values** instead of whole numbers, this suggests transactions could be **volume-based or weight-based**.

    - **Example:** A purchase of **2.5 kg of apples** or **1.75 liters of milk** rather than discrete units.
2. **FINAL_SALE Has Values Where FINAL_QUANTITY is 0**
    - Some transactions have `FINAL_QUANTITY = 0` but still have a **valid `FINAL_SALE` value**.
    - **Possible Reasons:**
        - Data entry or receipt scanning errors.
        - Items were refunded or adjusted post-purchase.
        - Promotional discounts or store-specific adjustments.
    - **Assumption:** These transactions were kept as they might still represent valid sales.
3. **Handling Missing BARCODE**
    - **Assumption:** Missing `BARCODE` values indicate unscanned or manually entered transactions.
4. **Date Consistency**
    - **Assumption:** `PURCHASE_DATE` is the actual transaction date, while `SCAN_DATE` represents when the receipt was uploaded.

5. Conclusion

- The **Transactions Dataset** required **data type corrections, duplicate removal, and assumption-based decisions** for missing values.
- **FINAL_QUANTITY was interpreted as both discrete and continuous (volume/weight-based) units**.
- **Some transactions had sales without quantity**, which were assumed valid.
- The dataset is now **clean and structured** for merging with the **Users and Products datasets** for further analysis.

I. Products Dataset Documentation

1. Dataset Overview

The **Products Dataset** consists of **845,552 records** and provides information on product details such as category, manufacturer, and brand. This dataset is essential for linking transaction data to product-level insights.

Columns:

Column Name	Description
CATEGORY_1	High-level product category (e.g., Snacks, Health & Wellness).
CATEGORY_2	More detailed sub-category within CATEGORY_1.
CATEGORY_3	Further refined classification of products.
CATEGORY_4	The most specific category level (significant missing values).
MANUFACTURER	The company that produces the product.
BRAND	The commercial name under which the product is sold.
BARCODE	Unique identifier for each product.

2. Data Quality Issues Identified

The following **data quality issues** were found during dataset exploration:

Missing Values

- **BARCODE**: **4,025 missing values** (~0.5% of the dataset).
- **BRAND**: **226,472 missing values** (~26.8% of the dataset).
- **MANUFACTURER**: **226,474 missing values** (~26.8% of the dataset).
- **CATEGORY_3** and **CATEGORY_4**: **Significant missing values** (CATEGORY_4 is missing in ~92% of records).

Data Type Issues

- **BARCODE was stored as `float64`**, which caused `.0` formatting issues when dealing with barcode numbers.

Placeholder Values

- `"Placeholder Manufacturer"` appeared **86,902 times**, indicating missing manufacturer information.
- `"Brand Not Known"` appeared **17,025 times**, indicating missing brand names.

3. Data Cleaning & Standardization

To address these issues, the following cleaning steps were applied:

Handling Missing Values

- **BARCODE**: Missing values were **left as NaN** since they cannot be inferred.
- **BRAND and MANUFACTURER**: Placeholder values (`"Placeholder Manufacturer"`, `"Brand Not Known"`) were replaced with **NaN** to avoid misleading analysis.
- **CATEGORY_3 and CATEGORY_4**: No imputation was done; missing values were retained. However, these columns **can be removed** as they were not essential for the analysis.

Data Type Conversions

- **BARCODE was converted to a string** to prevent floating-point formatting errors.

Duplicate Removal

- **400 exact duplicate records** were removed.

4. Assumptions & Considerations

During data cleaning and analysis, the following **assumptions** were made:

1. **BARCODE as the Primary Identifier**
    - Assumed that `BARCODE` uniquely identifies a product, even though **some barcodes were linked to multiple manufacturers or brands**.
    - Assumed that duplicate barcodes were valid due to product variations or licensing.
2. **Handling Placeholder Values**

    - `"Placeholder Manufacturer"` and `"Brand Not Known"` were assumed to be **missing values rather than actual categories**, and thus converted to `NaN`.
3. **Categorization Assumptions**
    - Since `CATEGORY_3` and `CATEGORY_4` had a **high percentage of missing values**, only `CATEGORY_1` and `CATEGORY_2` were considered reliable for analysis.
4. **Multiple Manufacturers and Brands per Barcode**
    - Some products had **multiple manufacturers or brands associated with the same barcode**.
    - Assumed that these variations were **valid** rather than data errors.
    - Instead of forcing a single record per barcode, **all variations were retained**.

5. Conclusion

- The **Products Dataset** required **duplicate removal, placeholder replacement, and type conversions**.
- **Missing values in key fields (BARCODE, BRAND, MANUFACTURER) were retained as NaN** instead of making assumptions.
- **Multiple manufacturers and brands per barcode were preserved**, assuming product variations are valid.
- The dataset is now **ready for merging with the Transactions Dataset** for product-level insights.

Fields That Are Challenging to Understand

1. FINAL_QUANTITY & FINAL_SALE in Transactions

Some records have **`FINAL_QUANTITY = 0` but a valid `FINAL_SALE` value**, which presents a challenge in determining the accuracy of these transactions. Possible explanations include:

- **Refunded items** where the sale remains recorded but the quantity is adjusted to zero.
- **Weight-based or volume-based purchases**, where `FINAL_QUANTITY` may not always reflect discrete units.
- **Data entry or system capture errors**, leading to inconsistencies in quantity reporting.

Challenge: These cases require assumptions on whether they represent valid transactions or data anomalies. For this analysis, these records were retained under the assumption that they might be system-recorded adjustments or valid non-unit-based sales.

2. **BARCODE in Products & Transactions**

Some barcodes are associated with **multiple manufacturers or brands**, leading to potential inconsistencies in product identification. Possible reasons include:

- **Product variations** where different manufacturers package the same barcode under private labels.
- **Licensing differences** where multiple companies produce the same item under agreements.
- **Duplicate or inconsistent data entry**, resulting in variations of manufacturer or brand details.

Challenge: It is unclear whether these differences are intentional (e.g., valid product variations) or data quality issues. For this analysis, all variations were retained, assuming they reflect genuine differences in product sourcing or branding.