**Data preparation:** prepare the data for analysis

1. Create MS SQL Server locally
2. Create a user database: takehome
3. Load 3 csv files into database takehome as table users, products and transactions

**Exercise:**

1. **Explore the data:**
   1. Data quality issues:
      1. Final\_quantity: should be numeric according to diagram, but this field has String in it
      2. Some of the products have null value in Barcode and length of barcode is different (3-14)
      3. Primary keys in tables are not unique, there are duplicated values in all 3 tables. The duplicated data need to be cleaned up and corrected (**Please refer to script Data\_Quality\_Check.sql**)
         1. Users: ‘id’ field
         2. Products: ‘barcode’ field
         3. Transactions: ‘receipt\_id’ field
      4. Majority of the user\_id in Transactions are missing from users table:

1. Select distinct

2.

3. user\_id,

4. id

5.

6. from takehome..transactions t

7. left join takehome..users u

8. on u.id = t.user\_id

9. where u.id is null

10. ;

* + 1. Some other fields have null values:
       1. Birth\_date
       2. State
       3. Gender
    2. Transaction table: Quantity is 0 but sale value is not 0, whereas quantity is not 0 but has sale value 0
  1. Are there any fields that are challenging to understand?

1. Language: es-419 is confusing
2. Category 1 – 4 are not well defined, users won’t know what the relationships are between these categories
3. Scan\_date vs purchase\_date: what scan\_date means?
4. What quantity means? The quantity that’s sold in this transaction or the quantity left after the transaction?
5. **provide SQL queries:** 
   1. Closed-ended Q1, What are the top 5 brands by receipts scanned among users 21 and over? **(Please refer to Close-Ended\_Q1.sql)**

|  |  |  |
| --- | --- | --- |
| **Brand** | **No\_of\_Receipts** | **Ranking** |
| DOVE | 6 | 1 |
| NERDS CANDY | 6 | 1 |
| SOUR PATCH KIDS | 4 | 3 |
| GREAT VALUE | 4 | 3 |
| COCA-COLA | 4 | 3 |
| HERSHEY'S | 4 | 3 |
| MEIJER | 4 | 3 |
| TRIDENT | 4 | 3 |

* 1. Closed-ended Q2, What are the top 5 brands by sales among users that have had their account for at least six months? **(Please refer to Close-Ended\_Q2.sql)**

|  |  |  |
| --- | --- | --- |
| **Brand** | **Accu\_Sales** | **Ranking** |
| CVS | 72.00 | 1 |
| TRIDENT | 46.72 | 2 |
| DOVE | 42.88 | 3 |
| COORS LIGHT | 34.96 | 4 |
| QUAKER | 16.60 | 5 |

* 1. Closed-ended Q3, What is the percentage of sales in the Health & Wellness category by generation? **(Please refer to Close-Ended\_Q3.sql)**

|  |  |  |  |
| --- | --- | --- | --- |
| **product\_category** | **user\_Gen** | **sale\_Gen** | **sale\_Percentage** |
| Health & Wellness | Senior | 75.12 | 47.09 |
| Health & Wellness | Old | 29.98 | 18.79 |
| Health & Wellness | Adult | 54.44 | 34.12 |

* 1. Open-ended Q2, Which is the leading brand in the Dips & Salsa category? **(Please refer to Open-Ended\_Q2.sql)**
     1. Approach 1: by # of receipts:

|  |  |  |
| --- | --- | --- |
| **Category\_2** | **Brand** | **# of Receipts** |
| Dips & Salsa | TOSTITOS | 56 |
| Dips & Salsa | PACE | 38 |
| Dips & Salsa | FRITOS | 31 |
| Dips & Salsa | MARKETSIDE | 23 |
| Dips & Salsa | DEAN'S DAIRY DIP | 22 |

* + 1. Approach 2: by sale amounts:

|  |  |  |
| --- | --- | --- |
| **Category\_2** | **Brand** | **Accu\_Sales** |
| Dips & Salsa | TOSTITOS | 181.3 |
| Dips & Salsa | GOOD FOODS | 94.91 |
| Dips & Salsa | PACE | 85.75 |
| Dips & Salsa | FRITOS | 67.16 |
| Dips & Salsa | MARKETSIDE | 65.22 |

1. **communicate email with stakeholders:**

Dear [Management Team],

I hope you are doing well. After analyzing the datasets, I have identified several **data quality issues** that may impact our ability to generate accurate insights. Additionally, there are some fields that require further clarification to ensure we interpret the data correctly. Below is a summary of the key findings:

**1. Data Quality Issues:**

* **Incorrect Data Types**: The Final\_quantity field should be numeric, but it contains text values, which may cause errors in calculations.
* **Missing & Inconsistent Barcode Data**: Some products have missing barcodes, and barcode lengths vary between 3 and 14 characters, which raises concerns about data consistency.
* **Duplicate Records**: The following primary keys contain duplicate values and need cleanup:
  + **Users Table**: id field
  + **Products Table**: barcode field
  + **Transactions Table**: receipt\_id field
* **User Data Mismatch**: A significant number of user\_id values in the **Transactions table** do not exist in the **Users table**, which suggests missing user records.
* **Null Values in Key Fields**: Several important fields contain missing values, including:
  + Birth\_date
  + State
  + Gender
* **Transaction Discrepancies**: There are cases where:
  + Quantity = 0 but **Sale Value is non-zero**
  + Quantity > 0 but **Sale Value is zero**  
    These inconsistencies may affect revenue calculations and need clarification.

**2. Unclear Fields & Definitions:**

* **Language Code (es-419)**: The meaning of this language code is unclear—does it refer to a specific region or setting?
* **Category Hierarchy**: Categories 1 – 4 are not well-defined, making it difficult to understand their relationships and relevance.
* **Scan Date vs. Purchase Date**: What does Scan\_date represent? Is it when the receipt was scanned, or when the transaction actually occurred?
* **Definition of Quantity**: Does Quantity refer to the number of items purchased in the transaction, or the remaining stock after the sale?

There is one interesting fact I found after analyzing the sales data by category is that about 66% of the sales in health and wellness category are coming from people older than 50, and none of the sales are coming from customers that are less than 30 as shown in below graph. In order to improve the sales number in this category, I suggest marketing team to target younger generation so that our customer base can be expanded easily.

**Next Steps & Support Needed:**

To resolve these issues and ensure reliable analysis, I would appreciate:

1. **Guidance on Handling Missing & Duplicate Data (especially duplicates on the primary key fields)** – Should missing values be estimated, ignored, or retrieved from another source?
2. **Clarification on Field Definitions** – A data dictionary or business context for ambiguous fields (es-419, category hierarchy, scan date vs. purchase date, quantity meaning).
3. **Confirmation on Transaction Inconsistencies** – Should zero-quantity sales be treated as errors, or is there a business reason behind them?

I’d be happy to discuss this further and propose solutions. Please let me know how you’d like to proceed.

Looking forward to your insights!

Best regards,  
Wenrui Zu  
Senior data analyst