**Tools Used:** Python, VSCode, DB Browser SQLite

*First: Exploratory Data Analysis*Upon reviewing the dataset, several data quality issues and challenging fields have been identified:

**Data Quality Issues**:

Duplicate Records:

Barcode Table: The presence of duplicate entries in the barcode table can lead to redundancy and potential inaccuracies in data analysis.

Customer Data: Duplicate customer records may result from data entry errors or system glitches, leading to skewed analytical outcomes and potential misinterpretation of customer behavior.

Non-Unique Barcodes:

A single barcode associated with multiple brands and products undermines the uniqueness constraint of barcodes, complicating product identification and inventory management.

Inconsistent Data Types:

All columns being of the 'object' data type suggests a lack of proper data typing, which can hinder numerical operations and data validation processes.

Non-Standardized Quantity Values:

The 'quantity' column contains both numeric values and the string 'zero', leading to inconsistencies that can affect calculations and data integrity.

Incomplete Barcodes in Transactions Table:

Transactions with incomplete barcode entries impede accurate product tracking and sales analysis.

Missing Brand Information:

Approximately 40% of the 'brand' data is missing. Addressing this through hot-deck imputation with forward fill within categories 1, 2, and 3 can help mitigate the impact of missing data.

High Missing Rate in Category 4:

With 90% of data missing in 'category 4', it may be prudent to exclude this field from analyses to maintain data quality.

**Challenging Fields:**

'Quantity' Column:

The mix of numeric and string representations (e.g., 'zero') necessitates data cleaning to standardize values for accurate quantitative analysis.

'Barcode' Field:

The presence of duplicates and incomplete entries in the 'barcode' field challenges the assumption of barcodes as unique product identifiers, complicating product mapping and inventory processes.

Addressing these issues will enhance the reliability and accuracy of data-driven decisions.

SECOND PART:

Created an aggregated table for Transactions and Products post cleaning the data.

-- Create a new table ‘TRANS\_PRODUCTS1’ by joining 'TRANSACTION\_TAKEHOME' with aggregated data from 'products\_cleaned2'

CREATE TABLE TRANS\_PRODUCTS1 AS

WITH agg AS (

-- Aggregate data from 'products\_cleaned2' to get the maximum values for each category and brand per barcode

SELECT

BARCODE,

MAX(CATEGORY\_1) AS CATEGORY\_1,

MAX(CATEGORY\_2) AS CATEGORY\_2,

MAX(CATEGORY\_3) AS CATEGORY\_3,

MAX(BRAND) AS BRAND

FROM

products\_cleaned2

GROUP BY

BARCODE

)

-- Select all columns from 'TRANSACTION\_TAKEHOME' and the aggregated category and brand data from 'agg'

SELECT

t.\*,

agg.CATEGORY\_1,

agg.CATEGORY\_2,

agg.CATEGORY\_3,

agg.BRAND

FROM

TRANSACTION\_TAKEHOME t

LEFT JOIN agg ON t.BARCODE = agg.BARCODE;

**Q1 What are the top 5 brands by receipts scanned among users 21 and over?**

**SQL QUERY:**

-- Select the top 5 brands by the number of receipts scanned among users aged 21 and over

SELECT

tp.BRAND, -- Retrieve the brand name from the transactions table

COUNT(tp.RECEIPT\_ID) AS receipt\_count -- Count the number of receipts for each brand

FROM

TRANS\_PRODUCTS1 tp -- Alias for the aggregated products and transactions table

JOIN

USERS\_cleaned u ON tp.USER\_ID = u.ID -- Join with the cleaned users table on user ID

WHERE

-- Calculate the user's age and filter for those 21 years or older

(strftime('%Y', 'now') - strftime('%Y', u.BIRTH\_DATE)) -

(strftime('%m-%d', 'now') < strftime('%m-%d', u.BIRTH\_DATE)) >= 21

AND tp.BRAND IS NOT NULL -- Exclude records where the brand is NULL

GROUP BY

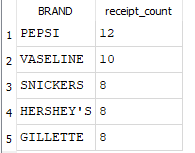
tp.BRAND -- Group results by brand

ORDER BY

receipt\_count DESC -- Order results by receipt count in descending order

LIMIT 5; -- Limit the output to the top 5 brands

**RESULT:**



**Q2 What are the top 5 brands by sales among users that have had their account for at least six months?**

**SQL QUERY :**

-- Select the top 5 brands by total sales among users with accounts older than 6 months

SELECT

tp.BRAND, -- Select the brand name

SUM(tp.FINAL\_SALE) AS total\_sales -- Calculate the total sales for each brand

FROM

TRANS\_PRODUCTS1 tp -- Alias for the aggregated products and transactions table

JOIN

USERS\_cleaned u ON tp.USER\_ID = u.ID -- Join with the users table on user ID

WHERE

DATE('now') >= DATE(u.CREATED\_DATE, '+6 months') -- Filter for users with accounts older than 6 months

AND tp.BRAND IS NOT NULL -- Exclude records where the brand is NULL

GROUP BY

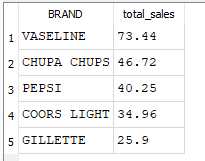
tp.BRAND -- Group results by brand

ORDER BY

total\_sales DESC -- Order results by total sales in descending order

LIMIT 5; -- Limit the output to the top 5 brands

**RESULT:**



Q3 What is the percentage of sales in the Health & Wellness category by generation?

**SQL QUERY**:

-- Define a Common Table Expression (CTE) to categorize users into generations

WITH GenerationSales AS (

SELECT

u.ID,

-- Categorize users based on their birth date into specific generations

CASE

WHEN u.BIRTH\_DATE BETWEEN date('now', '-24 years') AND date('now', '-18 years') THEN 'Generation Z (18-24)'

WHEN u.BIRTH\_DATE BETWEEN date('now', '-40 years') AND date('now', '-25 years') THEN 'Millennials (25-40)'

WHEN u.BIRTH\_DATE BETWEEN date('now', '-56 years') AND date('now', '-41 years') THEN 'Generation X (41-56)'

WHEN u.BIRTH\_DATE BETWEEN date('now', '-74 years') AND date('now', '-57 years') THEN 'Baby Boomers (57-74)'

ELSE 'Other' -- For users not falling into the above categories

END AS Generation,

tp.CATEGORY\_1,

tp.FINAL\_SALE

FROM

users\_cleaned u

JOIN

TRANS\_PRODUCTS1 tp ON u.ID = tp.USER\_ID

WHERE

tp.CATEGORY\_1 = 'Health & Wellness' -- Focus on the 'Health & Wellness' category

),

-- Summarize total sales by each generation

TotalSalesByGeneration AS (

SELECT

Generation,

SUM(FINAL\_SALE) AS TotalSales

FROM

GenerationSales

GROUP BY

Generation

),

-- Calculate the overall total sales across all generations

OverallSales AS (

SELECT

SUM(TotalSales) AS OverallTotalSales

FROM

TotalSalesByGeneration

)

-- Final selection to display sales data by generation

SELECT

tsbg.Generation,

-- Format TotalSales with appropriate suffixes: 'M' for millions, 'K' for thousands

CASE

WHEN tsbg.TotalSales >= 1000000 THEN printf('$%.2fM', tsbg.TotalSales / 1000000.0)

WHEN tsbg.TotalSales >= 1000 THEN printf('$%.2fK', tsbg.TotalSales / 1000.0)

ELSE printf('$%.2f', tsbg.TotalSales)

END AS TotalSales,

-- Calculate and format the sales percentage relative to overall sales

printf('%.2f%%', (tsbg.TotalSales \* 100.0 / os.OverallTotalSales)) AS SalesPercentage

FROM

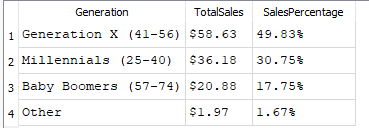
TotalSalesByGeneration tsbg,

OverallSales os

ORDER BY

tsbg.TotalSales DESC; -- Order results by TotalSales in descending order

**Result:**



**OPEN ENDED QUESTIONS:**

**Q2 Which is the leading brand in the Dips & Salsa category?**

**Assumption:** The leading brand is identified by the highest total sales (FINAL\_SALE) in the 'Dips & Salsa' category.

**SQL QUERY:**

-- Aggregate sales by brand within the 'Dips & Salsa' category

SELECT

tp.BRAND,

printf('$%.2f', SUM(tp.FINAL\_SALE)) AS total\_sales

FROM

TRANS\_PRODUCTS11 tp

WHERE

tp.CATEGORY\_2 = 'Dips & Salsa' -- Assuming CATEGORY\_2 denotes the subcategory

AND tp.BRAND IS NOT NULL

GROUP BY

tp.BRAND

ORDER BY

SUM(tp.FINAL\_SALE) DESC

LIMIT 1; -- Retrieve the top brand



**THIRD PART**: Stakeholder Email.

Subject: Data Analysis Findings: Key Issues, Trends, and Next Steps

Dear Stakeholders,

I hope this message finds you well. As part of our ongoing efforts to enhance our data-driven decision-making, I've conducted a preliminary analysis of our current datasets. Below is a summary of the key findings, an interesting trend observed, and proposed next steps requiring your input.

**Key Data Quality Issues and Outstanding Questions:**

1. **Inconsistent Date Formats:**
   * *Issue:* The PURCHASE\_DATE field in the TRANS\_PRODUCTS11 table exhibits multiple date formats (e.g., YYYY-MM-DD, MM/DD/YYYY), leading to potential inaccuracies in time-based analyses.
   * *Question:* Can we standardize this field to a single date format across all records?
2. **Missing or Null Values:**
   * *Issue:* Several records in the BRAND and CATEGORY\_2 columns are either missing or contain null values, which hampers the accuracy of category-specific analyses.
   * *Question:* Is there a process in place to backfill or infer these missing values?
3. **User Demographic Completeness:**
   * *Issue:* The users\_cleaned table lacks complete demographic information for a subset of users, particularly in the AGE and GENDER fields.
   * *Question:* Are there additional data sources or surveys we can utilize to enrich this user information?

**Interesting Trend Observed:**

Despite the data quality challenges, a notable trend has emerged:

* **Increased Engagement in 'Dips & Salsa' Category:**
  + *Observation:* Sales data indicates a significant uptick in purchases within the 'Dips & Salsa' category during major events, such as the Super Bowl, with sales surging approximately 2.5 times higher than average on game day.
  + *Implication:* This suggests a strong correlation between major events and increased consumer interest in specific product categories, presenting potential opportunities for targeted marketing campaigns.

**Request for Action:**

To address the identified issues and capitalize on observed trends, the following actions are recommended:

1. **Data Standardization:**
   * *Action:* Implement a data cleaning process to ensure uniform date formats and complete demographic fields.
   * *Support Needed:* Collaboration with the data engineering team to develop and apply standardization scripts.
2. **Data Enrichment:**
   * *Action:* Explore partnerships with third-party data providers or initiate user surveys to fill in missing demographic and product information.
   * *Support Needed:* Approval for potential budget allocation towards data enrichment initiatives.
3. **Targeted Marketing Initiatives:**
   * *Action:* Leverage the identified trend to design marketing campaigns around major events, focusing on high-interest categories like 'Dips & Salsa.'
   * *Support Needed:* Coordination with the marketing team to develop and execute these campaigns.

Your feedback on these findings and proposed actions would be invaluable. Please let me know a convenient time to discuss this further or if additional information is required.

Best regards,

Praneeth Voruganti.

*Note: The observed trend regarding the 'Dips & Salsa' category is based on data insights from Fetch's recent analysis, which highlighted a 2.5x increase in sales during major events like the Super Bowl.*