Fetch - Data Analyst Take Home

Exercise:

First: explore the data

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

- Are there any data quality issues present?
- Are there any fields that are challenging to understand?

Two step Approach

1. Data Ingestion : First step would be to import all three csv files as tables into a database lets call this database dbs. Let's assume our schema name is fetch , let's assume three new tables are created namely; fetch.product , fetch_user , fetch.transactions.
2. Data Exploration by each table : SQL Script to Check Data Quality Issues & Data Types

2.1) Products table : fetch.products , primary key barcode (underlying assumption , each barcode several columns have missing data, which may impact analysis and decision-making:

---

1. **Products.csv Data**

Data Quality Issues:

- CATEGORY_4: 92% missing (only 67,459 non-null out of 845,552).
- CATEGORY_3: 7.1% missing (784,986 non-null).
- MANUFACTURER & BRAND: 27% missing (~619,000 non-null).
- BARCODE: ~4,000 missing values (841,527 non-null).
- ***Potential Issue:*** *Missing CATEGORY_4 values might mean this field is optional or inconsistently filled. Missing MANUFACTURER and BRAND data could affect product identification and grouping.*
- **BARCODE Field Issue**: The BARCODE column is stored as float64, which is incorrect for a barcode.Floating-point storage may introduce rounding issues or convert long numeric values into scientific notation (e.g., 8.068109e+11).Some barcodes might have

non-numeric characters or inconsistent lengths.Potential Fix: Convert BARCODE to string (VARCHAR) in SQL to prevent rounding issues.

- **Duplicates & Data Integrity Issues** : If a duplicate check returns multiple records with the same BARCODE, CATEGORY_1, CATEGORY_2, etc., then:Either duplicates exist(bad data entry), OR Some fields are not unique enough to differentiate products.Potential Fix: Ensure BARCODE is a unique identifier if available. Otherwise, use a combination of fields (CATEGORY_1, BRAND, etc.) to detect true duplicates.
- **Inconsistent Category Hierarchy** : CATEGORY_1 → CATEGORY_2 → CATEGORY_3 → CATEGORY_4 should form a logical hierarchy.If a product has CATEGORY_3 but no CATEGORY_2, it violates the hierarchy.Potential Fix: Check if higher-level categories always exist when a lower-level one is present.

Challenging Fields:

- CATEGORY_4 Field : The high percentage of missing values makes it unclear whether this column is important.Does this field add value, or should it be ignored if CATEGORY_1-3 are enough?
- BARCODE as an Identifier:IIf BARCODE is missing or duplicated, what is the alternative unique identifier?Are manufacturer-assigned barcodes included, or are there

---

2. Transaction.csv Data

Data Quality Issues:

- Missing Values in BARCODE:
- Only 44,238 out of 50,000 transactions have a BARCODE.
- Some products might be missing barcodes, or there could be data entry errors.
- FINAL_QUANTITY & FINAL_SALE Have Inconsistent Values:
- FINAL_QUANTITY contains values like "zero", which should be numeric.
- Some FINAL_SALE entries are empty, possibly missing transaction amounts.
- Timestamp Format in SCAN_DATE & PURCHASE_DATE:
- Check for time zone mismatches.

Challenging Fields:

- FINAL_QUANTITY: Does "zero" mean a failed transaction, refund, or incorrect entry?
- FINAL_SALE: Missing values—are these zero-dollar transactions, or was the sale amount lost in processing?
- BARCODE Missing: How should transactions without barcodes be handled?private-label products without barcodes

SQL Script to Check Data Quality Issues & Data Types

This script will:

- Check data types for each column.
- Find missing values in each column.
- Identify duplicate records.
- Detect inconsistent data formats (e.g., non-numeric BARCODE values).
- Check string length variations for categorical column
- BARCODE Missing: How should transactions without barcodes be handled?

3.User.csv Data

Data Quality Issues:

- Missing Values:
  - BIRTH_DATE: Missing for 3.7% of users.
  - STATE: 4.8% missing.
  - LANGUAGE: Over 30% missing, making it unreliable for personalization.
  - GENDER: 5.9% missing, could be due to user preference or system error.
- Date Format Issues:
  - BIRTH_DATE and CREATED_DATE have timestamps- need to check with daat source provider if Birth Date needs stamp
  - Check for users without birth dates but valid accounts.

Challenging Fields:

- LANGUAGE: If 30% is missing, should it still be used for customer segmentation?
- GENDER: Should missing values be treated as "unknown", or is this a system flaw?
- BIRTH_DATE: Missing fields

- **Using SQL to identify Data Quality Issues taking products data as example , similarly this process can be repeated for user and transactions data.**

1. **Check Data Types of Each Column**

```
SELECT COLUMN_NAME, DATA_TYPE

FROM INFORMATION_SCHEMA.COLUMNS
```

```
WHERE TABLE_NAME = 'PRODUCTS';
```

2. **Check for Missing Values**

```
SELECT
   'CATEGORY_1' AS Column_Name, COUNT(*) AS Null_Count FROM PRODUCTS WHERE CATEGORY_1 IS NULL
UNION ALL
SELECT
   'CATEGORY_2', COUNT(*) FROM PRODUCTS WHERE CATEGORY_2 IS NULL
UNION ALL
SELECT
   'CATEGORY_3', COUNT(*) FROM PRODUCTS WHERE CATEGORY_3 IS NULL
UNION ALL
SELECT
   'CATEGORY_4', COUNT(*) FROM PRODUCTS WHERE CATEGORY_4 IS NULL
UNION ALL
SELECT
   'MANUFACTURER', COUNT(*) FROM PRODUCTS WHERE MANUFACTURER IS NULL
UNION ALL
SELECT
   'BRAND', COUNT(*) FROM PRODUCTS WHERE BRAND IS NULL
UNION ALL
SELECT
   'BARCODE', COUNT(*) FROM PRODUCTS WHERE BARCODE IS NULL;
```

3. **Check for Duplicate Records**

```
SELECT CATEGORY_1, CATEGORY_2, CATEGORY_3, CATEGORY_4, MANUFACTURER,
BRAND, BARCODE, COUNT(*)

FROM PRODUCTS

GROUP BY CATEGORY_1, CATEGORY_2, CATEGORY_3, CATEGORY_4,
MANUFACTURER, BRAND, BARCODE

HAVING COUNT(*) > 1;
```

4. **Check for Non-Numeric BARCODE Values**

```
SELECT BARCODE

FROM PRODUCTS

WHERE TRY_CAST(BARCODE AS BIGINT) IS NULL AND BARCODE IS NOT NULL;
```

5. **Detect String Length Issues in Categorical Columns**

```
SELECT

   'CATEGORY_1' AS Column_Name, MIN(LEN(CATEGORY_1)) AS Min_Length,
MAX(LEN(CATEGORY_1)) AS Max_Length FROM PRODUCTS

UNION ALL

SELECT

   'CATEGORY_2', MIN(LEN(CATEGORY_2)), MAX(LEN(CATEGORY_2)) FROM PRODUCTS

UNION ALL

SELECT

   'CATEGORY_3', MIN(LEN(CATEGORY_3)), MAX(LEN(CATEGORY_3)) FROM PRODUCTS

UNION ALL
```

```
SELECT

   'CATEGORY_4', MIN(LEN(CATEGORY_4)), MAX(LEN(CATEGORY_4)) FROM PRODUCTS

UNION ALL

SELECT

   'MANUFACTURER', MIN(LEN(MANUFACTURER)), MAX(LEN(MANUFACTURER)) FROM
PRODUCTS

UNION ALL

SELECT

   'BRAND', MIN(LEN(BRAND)), MAX(LEN(BRAND)) FROM PRODUCTS;
```

Second: provide SQL queries

Answer three of the following questions with at least one question coming from the closed-ended and one from the open-ended question set. Each question should be answered using one query.

Closed-ended questions:

- What are the top 5 brands by receipts scanned among users 21 and over

  ```
  —-#get all users of age 21 and above , where age is the difference in current year and birth date year. We will use a CTE for code efficiency and to avoid a user subquery on the main query down below.

With user_over_21 as

 (SELECT  distinct user_id

FROM om fetch.user

where birth_date is not null

AND date_diff('year' , birth_date :: date , getdate() :: date) >=21)
  ```

```
SELECT p.brand , count(distinct t.receipt_id) receipts_count

FROM fetch.transactions t

INNER JOIN fetch.products p on t.barcode = p.barcode —#left join to bring in brand from product table

Inner join user_helper_table uh on t.userid =uh.userid —-# only analyze for users above 21 years

WHERE p.brand is not null

Group by p.brand

Order by receipts_count desc

Limit 5;
```

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

  —get total sales from transactions by each brand, filter it on top 5 brands , further filter it for users who purchased at least 6 months prior to today's date

```
SELECT

   p.brand,

   SUM(t.final_sale) AS total_sales

FROM fetch.transactions t

JOIN fetch.product as p ON t.barcode=p.barcode

WHERE t.purchase_date <= DATEADD(month, -6, CURRENT_DATE)

GROUP BY

   p.brand

ORDER BY

   total_sales DESC

LIMIT 5;
```

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

  —classify user age into generations using birth year i am creating a 4 category based generation mapping to birth year

```
WITH UserGenerations AS (

   SELECT user_id,

Case when date_trunc('year' , birth_date') :: date BETWEEN '1946-01-01' AND '1964-01-01' then 'Baby Boomer'

        when date_trunc('year' , birth_date') :: date BETWEEN '1965-01-01'' AND '1980-01-01' then 'Gen X'

         When date_trunc('year' , birth_date') :: date BETWEEN '1981-01-01' AND '1996-01-01' then 'Millennial'

         When date_trunc('year' , birth_date') :: date '1997-01-01' AND '2012-01-01' then 'Gen Z'

         ELSE 'Alpha' END AS Generation  FROM fetch.user),

SalesByGeneration AS (

   SELECT U.Generation,

      SUM(case when p.CATEGORY_1 = 'Health & Wellness'  then CAST(T.FINAL_SALE AS FLOAT) ELSE 0 END) AS HealthWellnessSales,

      SUM(CAST(T.FINAL_SALE AS FLOAT)) AS TotalSales

   FROM fetch.transaction as T

   Inner join UserGenerations as U ON T.USER_ID = U.USER_ID

   Join fetch.products P ON T.BARCODE = P.BARCODE

   Where T.FINAL_SALE IS NOT NULL AND T.FINAL_SALE != ''

   GROUP BY U.Generation

)

SELECT
```

```
    Generation,
    (HealthWellnessSales * 100.0 / NULLIF(TotalSales, 0)) AS HealthWellnessPercentage
FROM SalesByGeneration;
```

Open-ended questions: for these, make assumptions and clearly state them when answering the question.

1. Who are Fetch's power users?

Assumptions:Definition of Power Users: Users who frequently scan receipts and have high total spending.Measurement Criteria:

- Number of transactions per user.
- Total sales per user.
- Frequency of purchases over time.

Top power users key metrics:

- Power users are those with the highest number of transactions and total sales.
- The top user (USER_ID: 64e62de5ca929250373e6cf5) has made 22 transactions, totaling $57.65 in sales.
- Another notable user (USER_ID: 60a5363facc00d347abadc8e) has fewer transactions (14) but the highest spending ($101.97).
- Conclusion: Fetch's power users are those who either:
- Frequently make purchases (high transaction count)
- Spend significantly on each transaction (high total sales)

2. Which is the leading brand in the Dips & Salsa category?

Assumption: We determine the leading brand by total sales revenue rather than the number of transactions or units sold.

- Finding: TOSTITOS is the top-selling brand in the Dips & Salsa category, with total sales of $103,354.84.

3. At what percent has Fetch grown year over year?

   Assumptions: Four assumptions - Sales Data is Complete and Accurate , Fetch's Business Model Remains Consistent , Excluding Invalid or Zero Sales Transactions are excluded, Yearly Sales Fluctuations Reflect Real Growth and Sales Growth is the Primary Metric of Business Growth.

```
WITH SalesByYear AS (
    SELECT left(purchase_date,4) as year,
        SUM(CAST(final_sale AS FLOAT)) AS Total_Sales
    FROM fetch.transactions
    Where FINAL_SALE IS NOT NULL AND FINAL_SALE != ''
    Group by year ),
YoY_Growth AS (
    SELECT
      year,
       Total_Sales,
       LAG(Total_Sales) OVER (ORDER BYyear) AS Previous_Year_Sales,
       ((Total_Sales - LAG(Total_Sales) OVER (ORDER BY Year)) * 100.0 /
        NULLIF(LAG(Total_Sales) OVER (ORDER BY Year), 0)) AS YoY_Growth_Percent
    FROM SalesByYear
)
SELECT * FROM YoY_Growth ORDER BY Year DESC;
```

Third: communicate with stakeholders

Construct an email or slack message that is understandable to a product or business leader who is not familiar with your day-to-day work. Summarize the results of your investigation. Include:

- Key data quality issues and outstanding questions about the data
- One interesting trend in the data
    - Use a finding from part 2 or come up with a new insight
- Request for action: explain what additional help, info, etc. you need to make sense of the data and resolve any outstanding issues

Email :

Hi Person X(Business/Product leader)

I've been reviewing Fetch's sales and user data, and wanted to share some key findings, along with a few areas where I need clarification to ensure we're making the most of our insights.

Insights and key trends:

- Our most active users fall into two groups:

1. Those who make frequent purchases (up to 22 transactions).
2. Those who spend more per transaction (one top user made only 14 purchases but spent over **$100**).

   This suggests we may have different types of high-value users—one driven by frequency, the other by order size. Understanding this split could help us tailor engagement strategies.

- Additionally, Tostitos is by far the best-selling brand in the Dips & Salsa category, with **$103K+** in sales—far ahead of competitors. It would be helpful to know whether this is due to special promotions, brand partnerships, or organic customer preference.

Next Steps & Support Needed

To refine these insights and ensure accuracy, I'd appreciate your inputs on:

1. **Handling missing/zero-dollar sales**—We found some zero dollar sales , could you please help share business insights on what type of sale is a zero dollar sale , this will help us decide if they should be excluded or investigated further?
2. **Purchase date inconsistencies**—To investigate if this is a known data issue, could you please direct me to the 'Purchase' team manager.
3. **Brand partnerships or promotions**—is there any additional context behind Tostitos' strong performance?

Would love your thoughts on these, and let me know if there's someone else I should check in with for further details. Happy to discuss!

Thanks and Regards,
Dakshika Chauhan