FETCH DATA ANALYSIS

Aditya Ramchandra Kutre
akutre@hawk.iit.edu

First: explore the data

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

1. Are there any data quality issues present?

2. Are there any fields that are challenging to understand?

Solution: Yes, there are data quality issues and challenging fields in each of the dataset given below is a detailed description for each file.

a.USERS_TAKEHOME.csv

Data Quality Issues:

- **Missing Values:**
  - **`BIRTH_DATE`** (3,675 missing values), **`STATE`** (4,812 missing values), **`LANGUAGE`** (30,508 missing values)**,** **`GENDER`** (5,892 missing values).
- **Data Consistency:**
  - **`BIRTH_DATE`** includes timezone information that might need standardization.
  - Some fields might contain inconsistent capitalization for `STATE` values.

Challenging Fields:

- `LANGUAGE`**:** Contains values like "es-419" which might not be immediately clear without understanding the encoding.(Needed only if the language field is being used for analysis).
- `BIRTH_DATE`**:** The format includes timestamps that may require additional parsing.
- `GENDER:` Having 12 unique values suggests non-standardized input or multiple gender classification systems .Without a data dictionary, it's unclear what all these gender options represent.

b.TRANSACTION_TAKEHOME.csv

Data Quality Issues:

- **Missing Values:**
  - `BARCODE` field has 5,762 missing values.
  - `FINAL_SALE` contains blank values, which may indicate missing financial data.
- **Data Format Issues:**
  - `FINAL_QUANTITY` has inconsistent entries such as "zero" instead of numeric values.
    - Unclear if fractional quantities (0.01) are valid or data errors.
    - No clear indication of the unit of measurement.
  - Dates (`PURCHASE_DATE`, `SCAN_DATE`) are in string format and may require conversion to proper datetime format.
  - `FINAL_SALE` has no currency indicator, further there is no clarity if the negative values are possible/valid.
  - Found 6 instances where `SCAN_DATE` is earlier than `PURCHASE_DATE`, which is logically impossible.

Challenging Fields:

- `FINAL_QUANTITY`**:** Inconsistent value types (e.g., "zero" instead of `0.00`).
- `SCAN_DATE`**:** Timestamp format with a time zone indicator may require additional processing to standardize.
- `BARCODE:` Different length barcodes present (some 11 digits, others 12) which is unusual since it is being mapped as primary key during mapping.

c. PRODUCTS_TAKEHOME.csv

Data Quality Issues:

- **Missing Values:**
  - `CATEGORY_1` (111 missing values), `CATEGORY_2` (1,424 missing values), `CATEGORY_3` (60,566 missing values), `CATEGORY_4` (778,093 missing values).
  - `MANUFACTURER` and `BRAND` have substantial missing data (~226K values missing).
  - `BARCODE` has 4,025 missing values.
- **Data Consistency:**
  - Due to multiple null values in each category there is an issue of inconsistent in depth of categorization across products.
  - The `BARCODE` field is in float format, which might introduce rounding issues and loss of precision.

Challenging Fields:

- `CATEGORY_3` and `CATEGORY_4`**:** Their hierarchical relationship with other category fields may not be clear without further documentation.
- `BARCODE`**:** Interpretation might be challenging due to its large numeric format stored as float, further the barcode length is inconsistent making it difficult to understand or map across the other dataset.

All the above conclusions have been drawn by using the dataset in python code attached. Further Pandas profiling has also been used to get a clear picture of each field helping to perform the Univariate and Bi/Multivariate analysis across each data column.

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:

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

Solution:

```
SELECT

P.BRAND, COUNT(T.RECEIPT_ID) AS total_receipts

FROM TRANSACTION_TAKEHOME T

JOIN USER_TAKEHOME U ON T.USER_ID = U.ID

JOIN PRODUCTS_TAKEHOME P ON T.BARCODE = P.BARCODE

WHERE TIMESTAMPDIFF(YEAR, U.BIRTH_DATE, CURDATE()) >= 21

GROUP BY P.BRAND

ORDER BY total_receipts DESC LIMIT 5;
```

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

```
SELECT
P.BRAND, SUM(T.FINAL_SALE) AS total_sales
FROM TRANSACTION_TAKEHOME T
JOIN USER_TAKEHOME U ON T.USER_ID = U.ID
JOIN PRODUCTS_TAKEHOME P ON T.BARCODE = P.BARCODE
WHERE TIMESTAMPDIFF(MONTH, U.CREATED_DATE, CURDATE()) >= 6
GROUP BY P.BRAND
ORDER BY total_sales DESC LIMIT 5;
```

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

```
SELECT
CASE
        WHEN TIMESTAMPDIFF(YEAR, U.BIRTH_DATE, CURDATE()) BETWEEN 18 AND 24 THEN 'Gen Z'
        WHEN TIMESTAMPDIFF(YEAR, U.BIRTH_DATE, CURDATE()) BETWEEN 25 AND 40 THEN 'Millennials'
        WHEN TIMESTAMPDIFF(YEAR, U.BIRTH_DATE, CURDATE()) BETWEEN 41 AND 56 THEN 'Gen X'
        WHEN TIMESTAMPDIFF(YEAR, U.BIRTH_DATE, CURDATE()) >= 57 THEN 'Boomers'
        ELSE 'Unknown' END AS generation,
SUM(T.FINAL_SALE) AS health_wellness_sales, (SUM(T.FINAL_SALE) / (SELECT SUM(FINAL_SALE) FROM TRANSACTION_TAKEHOME)) * 100 AS percentage_of_total_sales
FROM TRANSACTION_TAKEHOME T
JOIN USER_TAKEHOME U ON T.USER_ID = U.ID
JOIN PRODUCTS_TAKEHOME P ON T.BARCODE = P.BARCODE WHERE P.CATEGORY_1 = 'Health & Wellness'
GROUP BY generation
ORDER BY health_wellness_sales DESC;
```

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

a.Who are Fetch’s power users?

Assumption:

Power users are the one which make maximum transactions or the one which have made maximum sales.

Below query gives top 10 users with maximum transactions/spending.

SQL Query for determining the power users.

```
SELECT

   U.ID AS user_id,

   COUNT(T.RECEIPT_ID) AS total_transactions,

   SUM(T.FINAL_SALE) AS total_spent

FROM TRANSACTION_TAKEHOME T

JOIN USER_TAKEHOME U ON T.USER_ID = U.ID

GROUP BY U.ID

ORDER BY total_transactions DESC, total_spent DESC

LIMIT 10;
```

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

Assumption

Leading brand is the one which makes the highest sale in the category for **Dips & Salsa.**

SQL Query:

```
SELECT
   P.BRAND,
   SUM(T.FINAL_SALE) AS total_sales
FROM TRANSACTION_TAKEHOME T
JOIN PRODUCTS_TAKEHOME P ON T.BARCODE = P.BARCODE
WHERE P.CATEGORY_2 = 'Dips & Salsa'
GROUP BY P.BRAND
ORDER BY total_sales DESC
LIMIT 1;
```

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

Assumption:

Growth has been calculated with an assumption that data spans over multiple years with sufficient transactions being made.

SQL Query:

```
SELECT
    YEAR(PURCHASE_DATE) AS sales_year,
    SUM(FINAL_SALE) AS total_sales,
    LAG(SUM(FINAL_SALE)) OVER (ORDER BY YEAR(PURCHASE_DATE)) AS previous_year_sales,
    ((SUM(FINAL_SALE) - LAG(SUM(FINAL_SALE)) OVER (ORDER BY YEAR(PURCHASE_DATE))) /
      LAG(SUM(FINAL_SALE)) OVER (ORDER BY YEAR(PURCHASE_DATE))) * 100 AS year_over_year_growth
FROM TRANSACTION_TAKEHOME
GROUP BY YEAR(PURCHASE_DATE)
ORDER BY sales_year;
```

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:

Subject : Data Analysis Findings – Key Issues & Next Steps

Hi [Product/Business Leader's Name], Good morning!

Hope you are doing well!

I wanted to share some key findings from my recent analysis of our datasets and highlight a few areas where we may need additional support to ensure data accuracy and drive meaningful insights.

Key Data Quality Issues & Outstanding Questions:

1. **Missing Data:**
    - Significant gaps in critical fields such as `CATEGORY_3`, `CATEGORY_4`, and `MANUFACTURER` in the product data, which could impact our ability to analyze product-level performance accurately.
    - User demographic data, particularly `LANGUAGE` and `GENDER`, have missing values, which may limit our ability to segment and personalize user experiences effectively.
2. **Data Consistency:**
    - Inconsistent date formats across different tables (`CREATED_DATE` vs. `PURCHASE_DATE`), which might lead to challenges in tracking user activity over time.
    - Some `BARCODE` entries appear to be duplicated across different products, raising concerns about potential data integrity issues.

Would it be possible to investigate the data collection processes to identify and address the root causes of these inconsistencies?

Trend Insights Observed:

Based on the given dataset, I did go a step further and create a visualization using PowerBI, ignoring all the key data issues and fixing some of those such as format and duplicates.

We have identified that users who have had their accounts for over six months contribute to **a disproportionately high share of total sales**, suggesting that retention efforts could significantly boost revenue.

This insight highlights the importance of focusing on customer engagement strategies for long-term users to maximize their lifetime value.

Required Action:

To refine our analysis and address these challenges, I would need:

- Clarification on how missing product and user attributes are currently handled in our system.
- Additional context from the data engineering team to understand potential issues related to barcode duplication.
- Input from the marketing team to explore strategies for enhancing retention among long-term users.

Let me know if you have any questions or if we can set up a meeting to discuss further steps.

Please let me know if any additional information is required.

Thank you for your time. Have a great day ahead!

Thanks and Regards,

Aditya Ramchandra Kutre

akutre@hawk.iit.edu

Slack Message:

Hey @channel, 👋

I’ve been analyzing our recent datasets and wanted to share some key insights and next steps that I have found so far:

🚨 **Data Quality Issues:**

1. **Missing Data:** There are some critical fields like `CATEGORY_3`, `CATEGORY_4`, and `MANUFACTURER` are incomplete, which may impact product-level analysis.
2. **Inconsistencies:** Duplicate barcode entries and varying date formats across datasets could be affecting our reporting accuracy making the analysis more difficult.

Would appreciate input from the data engineering team to investigate these issues further.

📊 **Interesting Trend:**
Users who have been with us for over six months account for a **significant portion of total sales**, suggesting a strong case for targeted retention strategies.

📈 **Next Steps:**
I'm working on a Power BI dashboard to provide more actionable insights. Some planned visualizations include:

- **Sales Trends:** Year-over-year growth and category-wise performance.
- **User Behavior:** Age group analysis and top spending users.
- **Product Insights:** Best-selling brands and performance by store.

Let me know if you have any additional data requirements or insights you’d like to see included!

Thanks and Regards,
Aditya Ramchandra Kutre