Subject: Analysis of User Preferences and Data Quality: Actionable Insights

Hi,

After examining and conducting an analysis across the Users, Transactions, and Products datasets, I would like to share some important findings regarding data quality and related issues.

Key data quality issues

- **Insufficient Users Data -** The User_ID in the Transactions dataset does not match the ID in the Users dataset for approximately 99.48% of records. The significant mismatch limits the support in comprehensively analyzing through User demographic and targeting the Target Users/Audience of Fetch rewards.
  Recommendation: If we are focusing on transactions within the date range of June to September 2024, I recommend investigating the data sources to ensure that we are using a dataset that includes all users. This would enable seamless joins between the datasets and provide more robust statistical analysis for better insights.

- **Duplicated Records in Product DataSet** - There are duplicated entries in the Products table due to missing barcodes (NaN values). These duplicates may cause calculation errors when joining tables.
  Recommendation: Highly recommend that regularly updating and cleaning duplicate rows for effective data management, and benefits future analysis.

- **Account Creation Logic Issue -** There is one record showing a Create Date earlier than the User's Birth Date.
  Recommendation: To improve data quality, I recommend collaborating with product developers to implement validation logic that restricts invalid dates during account creation.

- **Transaction Create Logic Issue -** There are 94 records showing a Scan Date earlier than the Purchase Date. Based on my understanding, the Purchase Date refers to when the user purchases the items, and the Scan Date refers to when the user scans the receipt into the platform.
  Recommendation: I recommend implementing validation checks or adjusting the entry logic to ensure the Scan Date cannot be earlier than the Purchase Date. Additionally, it would be helpful to investigate whether the issue could be related to the scanning technology when reading the Purchase Date. Implementing these measures will help maintain data accuracy and prevent inconsistencies in transaction records.

- **Data Format Issue**

    - **FINAL_QUANTITY format -** The FINAL_QUANTITY field displays the text "zero" instead of the numeric value 0 in certain cases.
      Recommendation: Standardizing this field to always use numeric values (e.g., 0) is recommended for consistency in analysis and reporting.

    - **Date Format -** Most datetime columns are displayed in UTC format with a "Z" suffix, which is not user-friendly.
      Recommendation: I suggest changing the format to YYYY-MM-DD hh:mm:ss, aligning with what users typically expect.

    - **Barcode Format -** Barcodes are currently shown in scientific notation, although it is common for large numbers, scientific notation can cause issues in data analysis. It may lead to misinterpretation if analysts expect the barcode to be a string or a regular number.
      Recommendation: I also suggest changing the format to string or number aligning with what users typically expect.

- **Null data in datasets -** There are significant null values across multiple datasets, which can result in incomplete analysis or inaccurate results if not addressed. For example:
    - **Users Dataset:** 30.5% of the **Language** data is missing.
    - **Transactions Dataset:** 25% of the **FINAL_SALE** column is null.
    - **Products Dataset:** 26.78% of the **Manufacturer** and **Brand** columns are null.

  Recommendation: I recommend identifying the root causes of these null values and addressing them through data cleaning, imputation, or providing clearer guidelines for when null values should be expected.

- **Language is not clear -** Some language fields are inconsistent (e.g., using ‘es-419’ or ‘en’) which could cause confusion for analysts or stakeholders regarding the language or regional variations.
  Recommendation: I suggest adding a Language Description column to clarify the language and region, ensuring better understanding and reducing potential ambiguity in analyses.

Interesting Data Trending

Additionally, I'd like to highlight some interesting findings regarding user preferences. From my analysis of the Transactions and Production datasets from June to September 2024, it shows that Fetch's users show a strong preference for products in the Snacks, Health & Wellness, and

Beverages categories. Notably, **Carbonated Soft Drinks** dominate the **Beverages category**, accounting for **92.73% of all beverage transactions**.
To further validate these trends, I recommend testing these insights against historical data from the same months to determine if they are seasonal. Based on this analysis, targeting these preferences could help boost customer retention. For example, a promotional event like "**Bubble Summer: Earn More Points with Carbonated Soft Drinks**" could drive higher engagement and loyalty.

Please feel free to reach out if you have any questions.
Thank you!

Best,
Crystal Yu