Data Assessment Report

BY – AMOGH MAHADEV KOKARI (repo)

1. Key Data Quality Issues and Outstanding Questions

1.1 PRODUCTS_TAKEHOME.csv

- **Data Availability:** 78.08% of the data is available across the entire dataset.
  - **Missing Values:** The CATEGORY_4 column has the most missing values (778,093), while CATEGORY_1 has the least missing values (111).
- **Questions:**
  - Are the CATEGORY columns hierarchical or independent? They appear hierarchical, but a significant portion of the data is missing.
  - **Duplicates:** There are 185 duplicate BARCODE values.
  - **Barcode Assignment:** It seems that new BARCODES are assigned when some columns have missing data. Further clarification is needed on how BARCODE assignment works.

1.2 TRANSACTIONS_TAKEHOME.csv

- **Data Availability:** 95.43% of the data is available, with key columns such as RECEIPT_ID, PURCHASE_DATE, SCAN_DATE, STORE_NAME, USER_ID, and FINAL_QUANTITY having no missing values.
  - **Missing Values:** FINAL_SALE has the most missing values (12,500).
- **Data Quality Issues:**
  - FINAL_QUANTITY and FINAL_SALE contain string values ("zero" and " ").
  - **Negative Barcode Values:** The BARCODE column contains negative values.
  - **Outliers:** After cleaning the dataset, there are 64 outliers in FINAL_QUANTITY and 891 in FINAL_SALE (based on the assumption that outliers are any values beyond 2 standard deviations).
  - **Duplicates:** 171 duplicate rows were found.
  - **Date Issues:** The SCAN_DATE column contains timestamps, while PURCHASE_DATE only contains dates. Additionally, 94 rows have a PURCHASE_DATE that is later than the SCAN_DATE (potentially due to missing time components in PURCHASE_DATE).
  - **User-Receipt-Barcode Relationship:** Clarification is needed on how USER_ID, RECEIPT_ID, and BARCODE are related. Based on my understanding, one user can have multiple receipts, and one receipt can contain multiple barcodes, but barcodes are not unique to each receipt.
  - **Store Locations:** The store name appears to be the same even for chain stores. Why isn't location tracked?
  - **Units and Value Clarification:** How are FINAL_SALE and FINAL_QUANTITY units defined? Is FINAL_SALE the total dollar amount including taxes and discounts? Is FINAL_QUANTITY reported per unit or per package?

1.3 USERS_TAKEHOME.csv

- **Data Availability:** 92.51% of the data is available, with ID and CREATED_DATE having no missing values.
    - **Missing Values:** LANGUAGE has the most missing values (30,508).
- **Data Quality Issues:**
    - **Inconsistent Dates:** The user with ID 5f31fc048fa1e914d38d6952 has a BIRTH_DATE earlier than the CREATE_DATE.
    - **Underage Users:** 48 users registered on Fetch before turning 13 years old, with the youngest user being 2 years old.
    - **Overage Users:** 98 users registered on Fetch after turning 90 years old, with the oldest user being 125 years old. This may indicate potential fraudulent activity or a technical glitch.
    - **Clarification Needed:** What is the relationship between CREATE_DATE and BIRTH_DATE?
    - **Language Collection:** It appears that only two language options are available. Are these fixed?

2. Key Trends and Insights

2.1 User Growth Trend

- From 2014 to 2024, the user growth trend showed explosive growth initially, with an 820% increase in 2017, followed by steady gains through 2020. However, in 2021, the growth slowed to 13.48%, with a significant surge in 2022 (40% increase). In 2023, there was a sharp 42% decline, followed by a 25% drop in 2024.

  State-level Growth Trends:

    - **2020-2021:**
        - Best performing state: South Dakota (+60%)
        - Worst performing state: Alaska (-64.14%)
    - **2021-2022:**
        - Best performing state: Florida (+59.23%)
        - Worst performing state: Hawaii (-7.95%)
    - **2022-2023:**
        - Best performing state: Nebraska (-19.01%)
        - Worst performing state: Alaska (-70.91%)
    - **2023-2024:**
        - Best performing state: Montana (+22.22%)
        - Worst performing state: Nebraska (-54.08%)

Insight: The sharp drop in user growth from 2023 to 2024, especially in states like Alaska and Nebraska, requires further investigation to understand the underlying causes.

3. Request for Action: Additional Information Needed

To further refine this analysis and resolve outstanding questions, I require the following:

- **Data Structure Clarification:** A better understanding of the structure of the dataset, particularly the columns and their expected values. It would be helpful to have clarity on which columns should contain unique values and any assumptions about timestamps.
- **Business Context:** A deeper understanding of Fetch's business model and how it generates revenue. This will help connect the data to the company's strategic objectives and better understand the relevance of the collected data.
- **Insights for Stakeholders:** Information on the kinds of insights that are most valuable to stakeholders. Understanding what drives decision-making will help focus the analysis and ensure it aligns with business priorities.
- **Further Data Clarification:** Detailed descriptions of each column, including any constraints, data types, and expectations (e.g., acceptable ranges for FINAL_SALE and FINAL_QUANTITY).
- **Collaborative Discussion:** A discussion with other analysts or stakeholders to clarify any ambiguities, ensure all questions are addressed, and align the analysis with business objectives.