Deirdre Pethokoukis
Fetch Take Home Challenge
2/18/2025

EXPLORE THE DATA

Data exploration was performed on all three datasets. The SQL Script used to investigate the data is “02_FetchTakeHome_DataQualityChecks_DataIssuesOnly.sql”. In the following section, I will only highlight potential data issues that arose. A list of all data quality checks performed is found here “02a_FetchTakeHome_DataQualityChecks.xlsx” with the accompanying SQL Script “02a_FetchTakeHome_DataQualityChecks.sql”.

Additionally, a few graphs were made in Python for further understanding of the data. The Python Script used is “02b_FetchTakeHome_DataQualityChecks_Python.py”. Images of the graphs are in the folder “02b_FetchTakeHome_DataQualityChecks_Python_Graphs”.

TRANSACTIONS

I. Data Quality Issues
   - A. Row Counts
     1. The file contains exactly 50,000 rows. My first “gut feeling” is that these datasets are incomplete, and a sample was provided. The “Dataset Joins” section further details the issue of an incomplete dataset.
   - B. Unique Record
     1. There is no way to uniquely identify a record in the table. The table is either missing an additional field, duplication exists, or purchasing multiple units of the same product produces multiple lines (discussed more in “Duplicated Values” below).
   - C. Data Types
     1. PURCHASE_DATE has a datatype of “datetime” in the entity-relationship model. However, there is no time data. Therefore, to save space, it could be saved as a date.
     2. SCAN_DATE has a datatype of “datetime” in the entity-relationship model and contains time data. Therefore, it could be renamed to better reflect the time element (ex: SCAN_DATETIME).
   - D. Clean Values
     1. STORE_NAME has some values that could be cleaned to make values clearer/correct. For example, "IL''S WHIOSALE CLUB. Additional knowledge would be required to determine if the double apostrophes and misspellings are intentional.

2. The length of BARCODE is inconsistent. Most often, the length is 12. However, it can also have values of 2,8,9, and 13. Additional information is needed to know if this is normal.
3. FINAL_QUANTITY can have 'zero' instead of 0.00. This needs to be updated. Additionally, should this field have decimal values (i.e. 0.01) if it is a quantity?

E. Missing Values
1. BARCODE is missing for ~12% of records. What does it mean to have a missing BARCODE? Is this dataset incomplete or does it reflect an error in the scan? Given these rows can have FINAL_QUANTITY and FINAL_SALE, is it more likely the latter? This issue is prevalent across the entire date range of the dataset.
2. FINAL_SALE is missing for ~25% of records. This issue is prevalent across the entire date range of the dataset.

F. Duplicated Values
1. For a given RECEIPT_ID, the same BARCODE can have multiple entries. This is confusing because there is a FINAL_QUANTITY field. The FINAL_QUANTITY field should reflect instances where multiple units of the same product are purchased. Additional information on the FINAL_QUANTITY field would be helpful to determine if this is an issue.

G. Outliers
1. The range of FINAL_QUANTITY is 0.01 - 18.00. However, there is one value of 276. However, no barcode information in the PRODUCTS table to know if 276 units makes sense.
2. The range of FINAL_SALE is $0.00 to $139.31. However, there are 6 values of over $200. However, no barcode information in the PRODUCTS table to know if over $200 units makes sense. See the Python Graph "02_Final Sale Outliers". The data is very concentrated around zero. Therefore, additional analysis would be needed to understand outliers better.

H. Other
1. For a given RECEIPT_ID, it is not common to have multiple BARCODE values. Shouldn't a receipt usually have multiple products? Additional information on the BARCODE field would be helpful to determine if this is an issue.
2. ~25% of the time FINAL_QUANTITY is blank/null/zero and FINAL_SALE > 0. And ~26% of the time FINAL_QUANTITY > 0 and FINAL_SALE is blank/NULL/zero. How can a user buy no units but have

a price, or vice-versa? There is missing data or errors in the values picked up by the scan.

II. Fields Challenging to Understand
    A. It is unclear if the BARCODE is tied to the products being purchased as there is often only one BARCODE per RECEIPT_ID. Therefore, additional information on how this field is populated would be helpful.
    B. Additional information on FINAL_QUANTITY and FINAL_SALE as mentioned would be helpful as mentioned above. For example, it is unclear how they relate to each other as there are instances where only one is filled in.
    C. RECEIPT_ID (identifies unique receipt), PURCHASE_DATE (date products purchased), SCAN_DATE (date user scanned receipt), STORE_NAME (name of store found on receipt), and USER_ID (identifies unique user) make sense conceptually.

<u>USERS</u>

III. Data Quality Issues
    A. Row Counts
        1. The file contains exactly 100,000 rows. My first “gut feeling” is that these datasets are incomplete, and a sample was provided. The “Dataset Joins” section further details the issue of an incomplete dataset.
    B. Clean Values
        1. The values of GENDER are not standardized. For example, there is both "prefer_not_to_say" and "Prefer not to say". This field needs to be standardized.
    C. Missing Values
        1. CREATED_DATE does not have NULLs/blanks. However, there can be large gaps between dates when USERs were created. These gaps can suggest missing data.
        2. BIRTH_DATE is missing for 4% of records. It is missing in more recent CREATED_DATE years (2021+). Was this field made voluntary in 2021? Rows where BIRTH_DATE is missing are often missing other demographic info. There are also large gaps between when USERs were born. This is related to the issue of outliers mentioned below as the gaps primarily exist in very early or recent years.
        3. STATE is missing for 5% of records. LANGUAGE is missing for 31% of records. GENDER is missing for 6% of records. For all 3 fields, the issue is prevalent across the entire date range of the dataset. For STATE and LANGUAGE, it could be voluntary disclosure. Unless LANGUAGE is the language the app is presented in. However, since GENDER has

"prefer not to say" and "unknown" options, it is likely that there is still missing data.

D. Outliers
    1. The date range of BIRTH_DATE is 1/1/1900 and 4/3/2022. The pre-1920 and post-2010 dates seem the most unlikely. Additional information on how BIRTH_DATE is collected would be helpful in determining how erroneous these dates are.

IV. Fields Challenging to Understand
- A. Additional information on whether LANGUAGE is a voluntarily disclosed field or represents the language the app is used in would be helpful.
- B. All other columns make sense conceptually. However, additional information on which fields are voluntarily disclosed would help to understand which fields have missing data vs. omitted data.
    1. ID (identifies unique user), CREATED_DATE (when the user was created), BIRTH_DATE (user's birthdate), STATE (where the user lives), GENDER (user's chosen gender).

<u>PRODUCTS</u>

I. Data Quality Issues
- A. Unique Record
    1. There is no way to uniquely identify a record in the table. BARCODE is not a unique identifier due to the presence of NULLs in this field. Additionally, two BARCODEs have multiple entries, each with a different brand.
    2. Additionally, after converting BARCODE to an integer, duplication will exist if we join to the TRANSACTIONS table on BARCODE. The removal of leading zeroes leads to more BARCODEs having multiple entries with different MANUFACTURER/BRAND info.
- B. Clean Values
    1. CATEGORY_1 contains the value "Needs Review". These values need to be updated.
    2. MANUFACTURER has the value 'Placehold Manufacturer". These records need to be updated.
    3. The length of BARCODE is inconsistent. Most often, the length is 12. However, it can also have values of 6,7,8,9,10, 3, and 14. Additional information is needed to know if this is normal.
- C. Missing Values
    1. CATEGORY_1 is missing for <1% of records. The same manufacturers with "Needs Review" products have the missing categories. The categories get more specific with each number and the number of records

missing the category increases. This makes sense as products may not need additional specifications. Additionally, there is never an instance where CATEGORY_2/3/4 is not NULL and CATEGORY_1 is NULL. So at the minimum, all other products have CATEGORY_1 (outside the 111 where it is NULL). No investigation into the missingness of CATEGORY_2/3/4 was performed because of these notes.

2. MANUFACTURER is missing for 27% of records. BRAND is always missing in instances where MANUFACTURER is missing, except for 2 instances. BRAND is never missing when MANUFACTURER is filled in. The brand is "Listerine" in the two instances. For other products, Listerine has a manufacturer. It is unclear why it doesn't for these products.
3. BARCODE is missing for <1% of records. However, this is concerning as it can't connect the data back to the TRANSACTIONS data without BARCODE. This issue is prevalent across categories and manufacturers.

D. Outliers
   1. CATEGORY_1 values of "Health & Wellness" and "Snacks" are way higher in the count of products. Would need additional information on what products this dataset represents to know if it makes sense for these categories to be so large. See the Python Graph "04_Number of Records per Category 1".

E. Other
   1. It would be helpful to have a PRODUCT_NAME field to understand what each product is, rather than just a product category.

II. Fields Challenging to Understand
   A. Additional information on BARCODE would be helpful as described earlier.
   B. All the other columns make sense conceptually. CATEGORY_1 - CATEGORY_4 (describe the products with increasing specificity), MANUFACTURER (company who made the product), BRAND (product grouping product belongs to).

<u>DATASET JOINS</u>

The SQL Script "04_FetchTakeHome_CreateMasterTable.sql" details the issues when joining together the datasets into one master table. A data table with all percentages of matches can be found in "02a_FetchTakeHome_DataQualityChecks.xlsx".

I. Data Quality Issues
   A. When joining the TRANSACTIONS table to the USERS table using USER_ID and ID, very few matches are produced. Only 0.08% of the USER_IDs found in either table have a match in the other table.

B. When joining the TRANSACTIONS table to the PRODUCTS table on BARCODE, very few matches are produced. Only 0.78% of the BARCODEs found in either table have a match in the other table.
    1. Additionally, duplication is created as converting BARCODE to an integer (i.e. removing leading zeros) leads to more BARCODEs having multiple entries with different MANUFACTURER/BRAND info, as described above.
C. The lack of matches between the datasets points to all 3 datasets being incomplete. Because the TRANSACTIONS table is limited to 6/12/2024 - 9/8/2024, it makes sense that not all data found in USERS and PRODUCTS would be found in TRANSACTIONS. It wouldn't make sense for all users and all products that exist in these tables to be found in only 4 months.
D. However, it does not make sense for data found in TRANSACTIONS to be found in the other two lookup tables. Lookup tables should contain all users and products.

PROVIDE SQL QUERIES

The SQL Script "05_FetchTakeHome_SQLQueries.sql" contains all limitations, assumptions, and results for my chosen 3 questions.

The questions I chose were:

1. What is the percentage of sales in the Health & Wellness category by generation?
2. Who are Fetch's power users?
3. Which is the leading brand in the Dips & Salsa category?

COMMUNICATE WITH STAKEHOLDERS

An email was written for a business leader to summarize the results of my investigation. This email is below:

"

Subject: Fetch Data Review – Findings & Next Steps

Hi Business Leader -

I was able to review the Fetch datasets that provide transaction, product, and user data to assess data quality, identify interesting trends, and determine the next steps. See below for my findings:

Major Data Quality Issues and Questions

1. **Incomplete data** - Transactions only span 6/12/2022 - 9/8/2022. The product and user lookup table provided does not contain all the products and users found in the transactions data. Additionally, columns are missing as I am unable to identify what constitutes a unique record for the transactions and products datasets.
2. **Missing data** - In each of the 3 datasets, at least one field has missing values. For example, 92% of user data does not have the language field filled in. I need clarification on whether this is expected or an error.
3. **Outliers** - Some fields contain extreme values that need further investigation. For example, the birth date of some users is pre-1910.
4. **Incorrect Data** - There are fields requiring additional cleaning to standardize or correct data. For example, varying BARCODE lengths in the transactions/products data and GENDER in the users data has both “prefer_not_to_say” and “Prefer Not to Say” as options.

While reviewing the data, I found that “Walmart” is the most scanned store across all generations found in the data (from the Silent Generation to Generation Z). This suggests a strong Fetch presence for all ages at Walmart. However, this result is limited as I do not have the birthdates of most users and the data only spans 4 months. Therefore, this result is preliminary pending more complete data.

As for next steps, it would be helpful to:

1. Obtain a data dictionary for each dataset. This data dictionary would contain definitions of each field, how each field is populated, and expected values for each field. This would allow me to understand what values are valid and if missing data can or should be expected.
2. Confirm whether the datasets received are complete and if there are any additional fields not included in this production.

Once I receive more complete data, I will be able to use the data to produce more accurate results and findings. Please let me know who I should reach out to to discuss obtaining the data dictionary and more complete data.

Thank you,
Deirdre

”