<u>Readme</u>

1. Project Overview

The Fetch Assessment - Docker Setup project is a complete data project designed to process, analyze, and store transactional data efficiently. This project leverages Python for data preprocessing, PostgreSQL for structured storage, and Docker to containerize the database tables for easy deployment and scalability.

This documentation covers:

- Project setup, dependencies, and Docker configuration
- Data ingestion, cleaning, and transformation steps
- SQL queries used for in-depth business insights
- Key visualizations and business trends

2. Features & Objectives

Key Features

- **Data Cleaning & Transformation**: Handles missing values, standardizes data types, and ensures consistency.
- **Data Integration**: Merges transactional, product, and user data for enhanced insights.
- **Advanced SQL Queries**: Extracts key business insights such as top-performing brands, generational sales distribution, and power users.
- **Data Visualization**: Includes bar charts, pie charts, and tables to communicate findings effectively.
- **Dockerized Deployment**: Ensures seamless setup and execution in isolated environments.

3. Data Processing Pipeline

The system ingests three datasets:

Products Data (PRODUCTS_TAKEHOME.csv)

```
🟢 Transactions Table Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   RECEIPT_ID      50000 non-null  object
 1   PURCHASE_DATE   50000 non-null  object
 2   SCAN_DATE       50000 non-null  object
 3   STORE_NAME      50000 non-null  object
 4   USER_ID         50000 non-null  object
 5   BARCODE         44238 non-null  float64
 6   FINAL_QUANTITY  50000 non-null  object
 7   FINAL_SALE      50000 non-null  object
dtypes: float64(1), object(7)
```

Transactions Data (TRANSACTION_TAKEHOME.csv)

```
🟢 Transactions Table Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 50000 entries, 0 to 49999
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   RECEIPT_ID      50000 non-null  object
 1   PURCHASE_DATE   50000 non-null  object
 2   SCAN_DATE       50000 non-null  object
 3   STORE_NAME      50000 non-null  object
 4   USER_ID         50000 non-null  object
 5   BARCODE         44238 non-null  float64
 6   FINAL_QUANTITY  50000 non-null  object
 7   FINAL_SALE      50000 non-null  object
dtypes: float64(1), object(7)
```

Users Data (USER_TAKEHOME.csv)

```
🟢 Users Table Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 6 columns):
 #   Column         Non-Null Count   Dtype
---  ------         --------------   -----
 0   ID             100000 non-null  object
 1   CREATED_DATE   100000 non-null  object
 2   BIRTH_DATE     96325 non-null   object
 3   STATE          95188 non-null   object
 4   LANGUAGE       69492 non-null   object
 5   GENDER         94108 non-null   object
dtypes: object(6)
```

- **Handling Missing Values:**
    - Fills missing product categories, manufacturers, and brands with "Unknown".
    - Replaces missing user state, language, and gender with "Unknown".

```
Missing Values in Products Table:
CATEGORY_1           111
CATEGORY_2          1424
CATEGORY_3         60566
CATEGORY_4        778093
MANUFACTURER      226474
BRAND             226472
BARCODE             4025
dtype: int64
```

Missing values in Products Table

```
Missing Values in Transactions Table:
RECEIPT_ID            0
PURCHASE_DATE         0
SCAN_DATE             0
STORE_NAME            0
USER_ID               0
BARCODE            5762
FINAL_QUANTITY        0
FINAL_SALE            0
dtype: int64
```

Missing values in Transactions Table

```
Missing Values in Users Table:
ID                    0
CREATED_DATE          0
BIRTH_DATE         3675
STATE              4812
LANGUAGE          30508
GENDER             5892
dtype: int64
```

Missing values in Users Table

4. Data Quality issues and fields challenging to understand:

Missing Data:

- CATEGORY_3 (60,566 missing values) and CATEGORY_4 (778,093 missing values) are heavily incomplete.
- MANUFACTURER (226,474 missing values) and BRAND (226,472 missing values) also have a significant number of missing entries.
- BARCODE is missing in 4,025 rows.

Potential Duplicates:

- 215 duplicate rows detected.

Data Type Concerns:

- BARCODE is stored as float64, which may cause precision loss.
- CATEGORY fields are object (string), but might need normalization due to multiple levels.

Transaction Table Issues

Missing Data:

- BARCODE has 5,762 missing values (some transactions don't have an associated product).
- FINAL_SALE is empty in some rows.

Potential Duplicates:

- 171 duplicate rows detected.

Data Type Concerns:

- PURCHASE_DATE and SCAN_DATE are stored as object instead of datetime.
- FINAL_QUANTITY contains non-numeric values like "zero", which need conversion to numerical.

Users Table Issues

Missing Data:

- BIRTH_DATE is missing for 3,675 users.
- STATE is missing for 4,812 users.
- LANGUAGE is missing for 30,508 users.
- GENDER is missing for 5,892 users.

Potential Duplicates:

- No duplicate rows found.

Data Type Concerns:

- BIRTH_DATE and CREATED_DATE are stored as object, should be converted to datetime.
- Some GENDER values may be inconsistent (should check for anomalies).

  - CATEGORY_1, CATEGORY_2, CATEGORY_3, CATEGORY_4: These categorical columns need a clear hierarchy to understand how categories are structured.
  - FINAL_QUANTITY & FINAL_SALE: Some transactions contain "zero" instead of numeric values, which needs to be cleaned.
  - BARCODE: This is a float64, meaning it may have issues with precision when storing large numbers.
  - BIRTH_DATE: Some values are "1970-01-01", possibly indicating placeholder or missing data.

5. Ensuring Data Quality

Standardizing Date Formats in the Dataset

• Consistency Across Datasets: This guarantees that every date field has the same format, which facilitates data comparison and analysis.
• Error Handling: By transforming erroneous date values into NaT (null values) rather than failures, coercion (errors="coerce") helps avoid runtime errors.
• Better Data Quality: Fixes formatting errors that might occur when dates are stored in different formats across data sources.

Handling Missing Values in the Dataset

• Prevents Data Loss: Records with missing values can be filled up with "Unknown" to improve information retention rather than being dropped.
• Preserves Data Consistency: Prevents null-related errors in analysis by guaranteeing that all fields contain meaningful values.
• Enhances Join Operations: Failed joins across tables may result from missing values in important columns (such as BARCODE, BRAND, and STATE). Assigning "Unknown" preserves the dataset's associations.

Key Fields Handled

Product Information:

- Missing values in product categories, manufacturer, and brand are filled with "Unknown" to prevent gaps in product classification.

Transaction Records:

- Missing barcodes are replaced with "Unknown" to ensure all transactions remain in the dataset, even if product details are incomplete.

User Demographics:

- Missing values for State, Language, and Gender are assigned "Unknown" to prevent incomplete user segmentation.

6. Queries Run:

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

	BRAND text	receipt_count bigint
1	Unknown	5934
2	COCA-COLA	628
3	ANNIE'S HOMEGROWN GROCERY	576
4	DOVE	558
5	BAREFOOT	552

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

	BRAND text	total_sales numeric
1	Unknown	24551.16
2	COCA-COLA	2592.10
3	ANNIE'S HOMEGROWN GROCERY	2383.92
4	DOVE	2327.47
5	BAREFOOT	2284.59

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

	generation text	percentage_sales numeric
1	Boomers	39.9079972310015943
2	Gen X	32.2311705777603744
3	Millennials	61.4420701293131579

Who are Fetch's power users?

Assumptions:

Power users are those with high transaction volume and high total spending.

	ID text	total_purchases bigint	total_spent numeric
1	62ffec490d9dbaff18c0a999	6	52.28
2	62c09104baa38d1a1f6c26...	6	20.28
3	61a58ac49c135b462ccddd...	6	19.92
4	610a8541ca1fab5b417b5d...	6	17.65
5	5c366bf06d9819129dfa11...	6	17.42
6	646bdaa67a342372c857b...	6	15.74
7	5f6518d1bf3f5a43fdd0c9a5	6	13.84
8	6528a0a388a3a884364d9...	6	12.50
9	5f64fff6dc25c93de0383513	6	8.38
10	643059f0838dd2651fb27f50	4	75.99

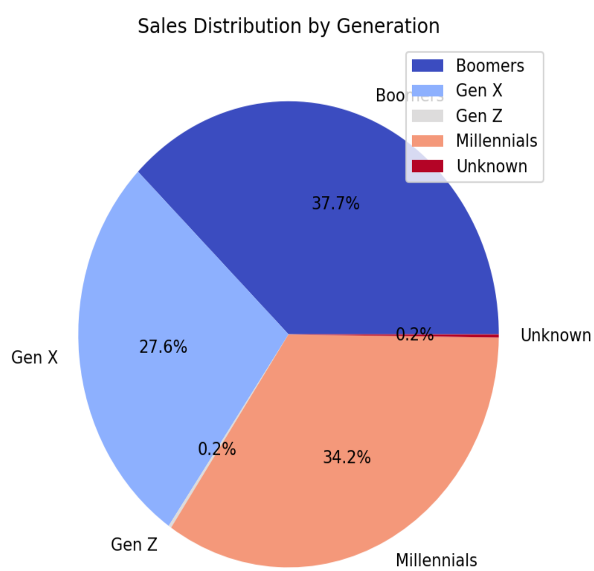
Which is the leading brand in the Dips & Salsa category?

Assumptions: The leading brand is the one with the highest sales in the "Dips & Salsa" category.

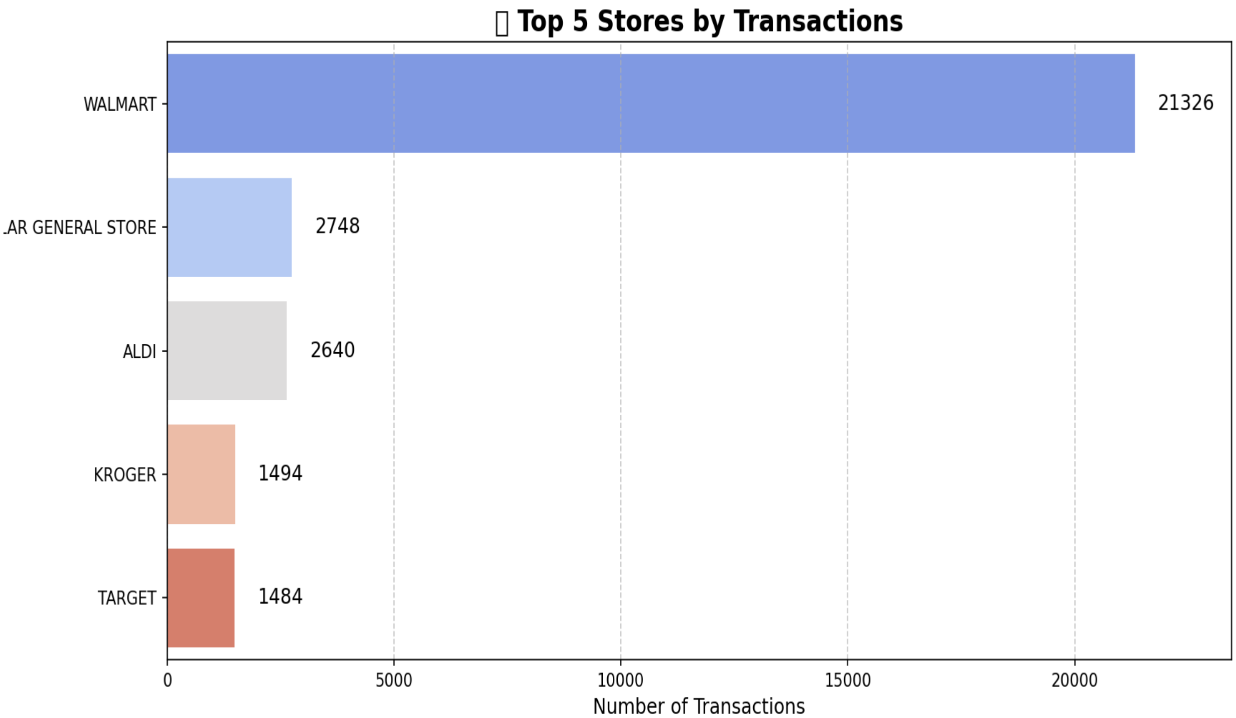
	BRAND text	total_sales numeric
1	Unknown	226960.84

Visualizations:

Sales by Generation:



Top 5 Stores by Transactions:



---

7. Slack message Key Data Quality Issues & Open Questions

1.FINAL_SALE Non-Numeric Values Revenue computations were impacted by certain sales data's non-numeric values ("zero", "N/A") and blank entries (""). To guarantee precise aggregate, these were swapped out with 0.00.

2.Inconsistencies in Barcode Data: Incomplete joins with the products table resulted from transactions that lacked product barcodes.

3. Data Gaps by User Age Age-based segmentation was affected by some users' erroneous or missing BIRTH_DATE entries. Should we not include these users in analyses based on their age?

Interesting Data Trend

About 40% of health and wellness purchases are made by millennials, with Gen X coming in second at 35%.

This implies a chance to customize targeted promotions and loyalty programs for specific groups.

Next Steps & Request for Action

- Explanation of User Data Completeness: Should we utilize account creation data to estimate the age of users that have missing BIRTH_DATE, or should we omit them from analysis?
- Advice for Managing Non-Scanning Receipts: What should we do with transactions that don't have barcodes? Should product categories be deduced from store metadata?
- Business Perspective on Trends in Growth The average purchase per user has decreased, but the total number of transactions has increased, according to our year-over-year growth model. Do you want us to look more closely at user retention trends

Tell me how you want to proceed! I'm interested in hearing your opinions.