

Fetch Data Analyst Take-Home

Understanding the Datasets

The provided datasets give us a detailed view of Fetch users, their purchase transactions, and the products they buy. Each dataset plays a crucial role in analyzing user behavior, sales trends, and brand performance:

- **Users Dataset:** Contains demographic details like USER_ID, AGE, GENERATION, SIGNUP_DATE, and STATE. This helps in understanding user demographics, engagement levels, and retention patterns.
- **Transactions Dataset:** Tracks purchase activity, including TRANSACTION_ID, USER_ID, RECEIPT_DATE, SALES_AMOUNT, CATEGORY_1, and CATEGORY_2. This data is essential for sales performance analysis and consumer behavior insights.
- **Products Dataset:** Provides details about the items purchased, such as PRODUCT_ID, BRAND, CATEGORY_2, and BARCODE. This dataset helps in identifying top brands and product trends.

Data Quality Issues

While exploring the data, I found some key quality issues that could impact analysis accuracy. These need to be addressed to ensure meaningful insights.

1. Missing Data

- Some products are missing values in CATEGORY_1, CATEGORY_2, and BARCODE, which can make it difficult to correctly categorize and track products.
- The STATE field is missing for some users, limiting our ability to perform location-based analysis (e.g., regional trends, purchase behaviors by state).
- SALES_AMOUNT is missing in some transactions, which can distort revenue calculations and sales performance insights.

2. Duplicate Records

- The PRODUCTS_TAKEHOME.csv dataset contains duplicate barcodes, which may indicate incorrect product mapping. This could lead to inflated sales counts or incorrect brand performance reports.

Fields That Are challenging to Interpret

- **GENERATION Classification:** The GENERATION field groups users by age, but it's unclear how these categories are defined. Are Millennials and Gen Z classified using standard industry definitions? Clarifying this would help ensure consistency in demographic insights.
- **CATEGORY_2 in Transactions:** Some transactions don't have a CATEGORY_2 value, making it harder to analyze product subcategories. If this information isn't consistently captured, it could affect category-specific sales analysis.
- **BARCODE in Products:** Some products have missing BARCODE, making it difficult to track unique items.