**Key Findings**

**Are there any data quality issues present?**

Yes, there are several data quality issues observed across the three datasets: PRODUCTS\_TAKEHOME, TRANSACTION\_TAKEHOME, and USER\_TAKEHOME. Starting with the PRODUCTS\_TAKEHOME csv, I found missing values in several critical columns including CATEGORY\_1, CATEGORY\_2, CATEGORY\_3, CATEGORY\_4, MANUFACTURER, BRAND, and BARCODE. These fields are essential for uniquely identifying and categorizing products and missing data here could impact analysis involving product segmentation or brand- trends forecasting.

Similarly, the TRANSACTION\_TAKEHOME csv showed missing BARCODE entries. Since this field is essential for linking a transaction to a product, its absence can severely limit our ability to perform joined analyses across transactions and products. Additionally, I observed potential issues in numeric fields like FINAL\_SALE and FINAL\_QUANTITY that would need validation to ensure no negative or zero values where inappropriate.

In the USER\_TAKEHOME csv, fields like BIRTH\_DATE, STATE, LANGUAGE, and GENDER had missing entries. Additionally, some date fields like CREATED\_DATE and BIRTH\_DATE had invalid or corrupt entries that got converted to NaT (Not a Time) upon parsing. These are indicative of formatting issues or corrupt values in the original CSVs. We also identified duplicate rows, which point to potential redundancy or improper record-keeping.

I assume the missing values are due to input errors or incomplete system logs

From concluding that this dataset has common data quality issues such as missing values, duplicates, and unparseable dates. These will need cleaning and imputation strategies before any analysis or modeling.

**Are there any fields that are challenging to understand?**

Yes, a few fields across the datasets appear unclear or lack proper situation, making them harder to understand directly.

In the PRODUCTS\_TAKEHOME dataset, fields such as CATEGORY\_1, CATEGORY\_2seem to represent a hierarchy of product categories. However, there’s no additional data dictionary or description to help us understand what each category level signifies. For example, does CATEGORY\_1 represent the department like ‘Beverages’, while CATEGORY\_4 is a more specific subcategory like ‘Diet Soda’. Without additional metadata, it's difficult to infer the relationships between these category columns.

In the TRANSACTION\_TAKEHOME file, while most columns are straightforward, the distinction between PURCHASE\_DATE and SCAN\_DATE could be confusing. Do they represent the actual date of transaction and the date it was logged into the system respectively? If so, the channel between these fields might indicate system delays or offline purchases. Similarly, the field FINAL\_SALE needs clarification is it the gross sale amount, or is it net of discounts and tax?

For the USER\_TAKEHOME dataset, the CREATED\_DATE field presumably refers to the date the user account was created, but without knowing how users interact with the platform or system, its significance is limited. Furthermore, some categorical fields such as LANGUAGE and STATE have missing or inconsistent values. The GENDER field, although common, might also contain non-standard entries depending on how the data was collected.

To make these fields more understandable, either data documentation or EDA helps to understand better and would bring necessary results.