## Data Exploration

During the data exploration I came across the following issues:

### 1. Missing Values

Below are some of them:

| Table | Column | # Missing Values | % Missing values | Comments |
|---|---|---|---|---|
| PRODUCTS_TAKEHOME | CATEGORY_1 | 111 | 0.013% | Category_1 is the top level of the hierarchy, and should probably have a value, since each product could be classified as belonging to some category. |
| PRODUCTS_TAKEHOME | CATEGORY_4 | 778,093 | 92% | Very high number, suggesting it may not be essential or poorly maintained |
| PRODUCTS_TAKEHOME | MANUFACTURER | 226,474 | 26% | 26% have missing **MANUFACTURER and BRAND** |
| PRODUCTS_TAKEHOME | BRAND | 226,472 | 26% | |
| PRODUCTS_TAKEHOME | BARCODE | 4,025 | 0.48% | |
| TRANSACTION_TAKEHOME | BARCODE | 5,762 | 11.52% | |
| TRANSACTION_TAKEHOME | FINAL_SALE | 12,500 | 25% | |
| USERS_TAKEHOME | GENDER | 5,892 | 5% | |
| USERS_TAKEHOME | STATE | 4,812 | 4% | The missing percentage is not high, however state can be important upon |

| | | | | |
|---|---|---|---|---|
| | | | | transaction for tax calculations |
| USERS_TAKE HOME | BIRTH_DATE | 3,675 | 10% | Date of birth could be important to determine if user is not a minor and eligible for purchasing products |

- Additional columns contain missing values, but that might be valid. For instance:
  **CATEGORY_2 & CATEGORY_3** have missing values as well, some products might not have all hierarchy levels populated.
  **Language** has about 30% missing values, the high volume of missing values indicates that perhaps this was a non mandatory field when users signed up

## 2. Duplications

- **PRODUCTS_TAKEHOME** - **215** duplicated rows
- **TRANSACTION_TAKEHOME** - **171** duplicated rows

## 3. Data Join Issues: Missing Matches

- When joining **PRODUCTS_TAKEHOME** with **TRANSACTION_TAKEHOME 50%** of the transactions can't be matched with a product
- When joining **TRANSACTION_TAKEHOME** with **USERS_TAKEHOME 99%** of the transactions can't be matched with a user

## 4. Temporary Data Values: Potential Considerations

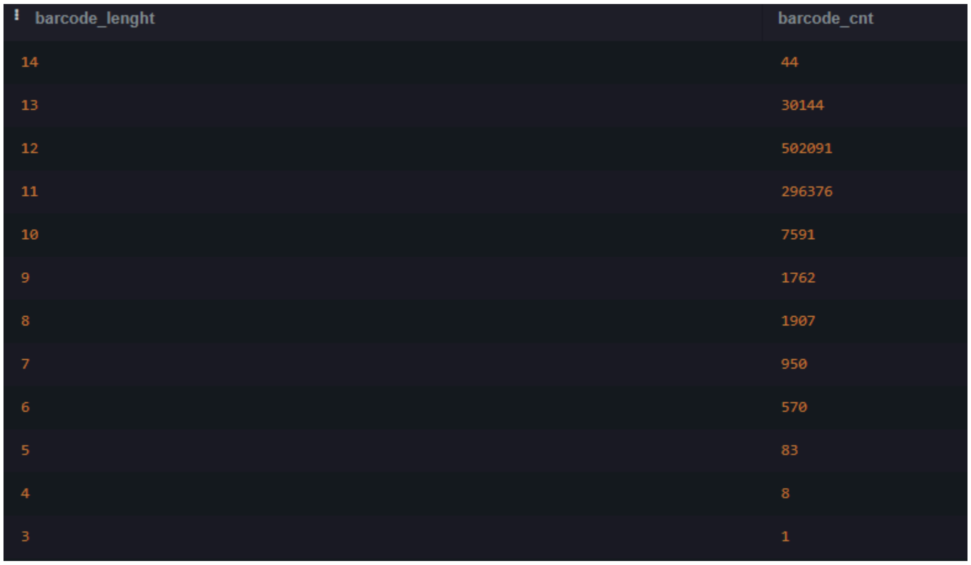
These values may not be an issue as long as they are accounted for in analysis and BI processes, and relevant stakeholders are aware.

- **PRODUCTS_TAKEHOME** Table: 86,902 rows have "PLACEHOLDER MANUFACTURER" in the MANUFACTURER field.
- **PRODUCTS_TAKEHOME** Table: 547 rows have "Needs Review" in the CATEGORY_1 field.
- **PRODUCTS_TAKEHOME** Table: 17,025 rows (2%) have "BRAND NOT KNOWN" or "BRAND NEEDS REVIEW" in the BRAND field.

## 5. Barcode Values

- **PRODUCTS_TAKEHOME & TRANSACTION_TAKEHOME** - Barcodes are typically 12 or 13 digits, with some exceptions for specific products. The data set contains different lengths, this could indicate corrupted data. Here is the barcode count per length from

PRODUCTS_TAKEHOME



| barcode_lenght | barcode_cnt |
|---|---|
| 14 | 44 |
| 13 | 30144 |
| 12 | 502091 |
| 11 | 296376 |
| 10 | 7591 |
| 9 | 1762 |
| 8 | 1907 |
| 7 | 950 |
| 6 | 570 |
| 5 | 83 |
| 4 | 8 |
| 3 | 1 |

- **PRODUCTS_TAKEHOME** - 54 barcodes are not unique, but the products are not duplicated. Meaning, rows with the same barcode having differences in other columns (e.g. different brand).
- **TRANSACTION_TAKEHOME** - a single row contains 'BARCODE' as a value, 8 rows contain '-1' as the barcode value

## 6. FINAL_SALE Values

**TRANSACTION_TAKEHOME** - some FINAL_SALE values are null, even when FINAL_QUANTITY is > 0 or when the same receipt has another record with non null FINAL_SALE value

## 7. FINAL_SALE Values

**TRANSACTION_TAKEHOME -** FINAL_QUANTITY contains 'zero' a non numeric value. This would be problematic when conducting a numeric analysis

## 8. Big variety of categorical values

This could make analysis and grouping difficult:

- **USER_TAKEHOME** - **Gender -** has **11** unique values. The values aren't standardized - some have the same meaning with different spelling- non_binary & Non-Binary, prefer_not_to_say & Prefer not to say, not_listed & My gender isn't listed.
- **PRODUCTS_TAKEHOME** - **MANUFACTURER** and **BRAND** have a high count of distinct values (4,354 and 8,122 distinct values)

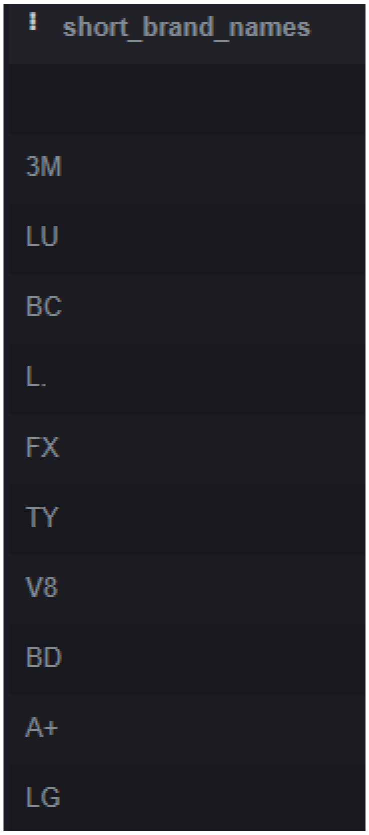
## 9. Poor Naming

**PRODUCTS_TAKEHOME** - CATEGORY_1, CATEGORY_2, CATEGORY_3,CATEGORY_4 are not descriptive

- **TRANSACTION_TAKEHOME** - FINAL_SALE field is not a self explanatory name, if it’s intended to represent paid amount it should explicitly say this.

## 10. Brand Names

**PRODUCTS_TAKEHOME** - Some brands have short names. Some are valid, such as LG, but when googling I couldn’t find others



| short_brand_names |
|---|
| |
| 3M |
| LU |
| BC |
| L. |
| FX |
| TY |
| V8 |
| BD |
| A+ |
| LG |

## 11. Data Type

- **USER_TAKEHOME** - BIRTH_DATE and CREATED_DATE should be of daytime type instead of  textual
- **TRANSACTION_TAKEHOME** - PURCHASE_DATE and SCAN_DATE are both in text format, and each of them follows different conventions (date only vs. date + time). It’s bad practice to use text format for datetime fields, and both fields should follow the same convention.

## 12. Other Outliers

**USER_TAKEHOME** - User with the ID 5f31fc048fa1e914d38d6952 has create_date > birth_date