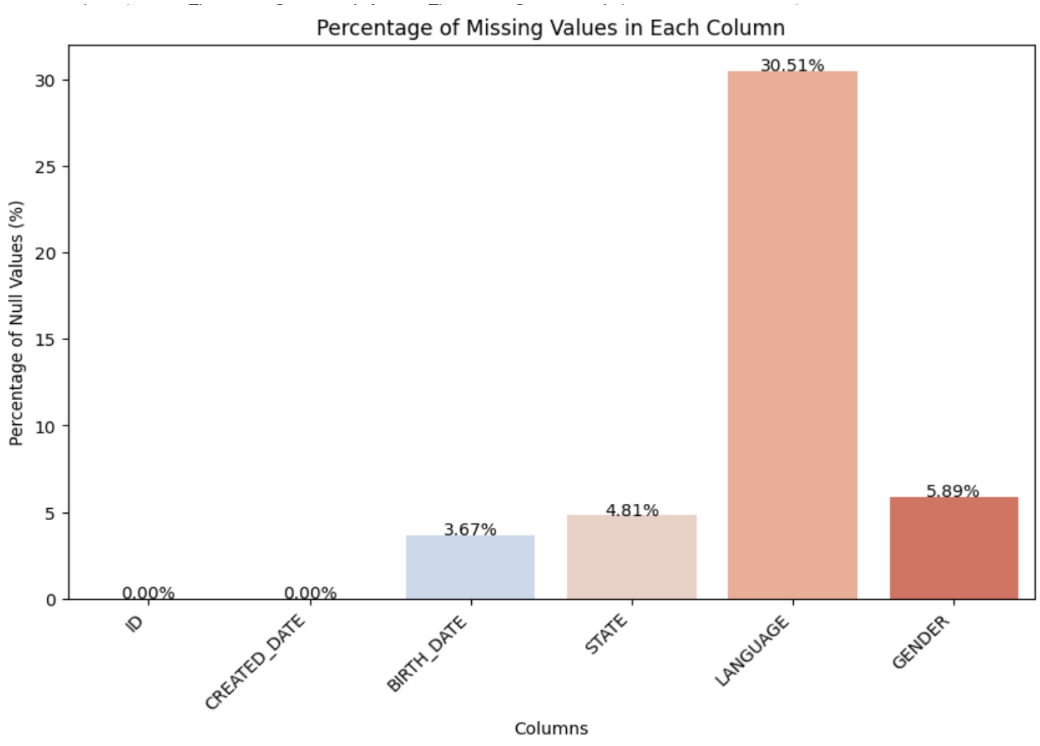
Description of Given Data

Users:

The **Users** table provides demographic information about users in the Fetch application, including **gender, state, date of birth, user ID, language, and account creation date**. The absence of null values in the **"Created Date"** column confirms that this dataset pertains exclusively to registered Fetch users.

Data Quality Issues:

• The dataset contained a significant number of **null or missing values** across most columns, with the exception of UserID and Created Date. Out of 10,000 records in the table, following is the null distribution.



Rather than deleting the null values, which could result in the loss of valuable user and transaction data, I opted to normalize the null values. This approach was taken to ensure better understanding and facilitate more accurate analysis.

• There were **inconsistencies in the data types** as well. The **Created Date** and **Birth Date** columns, which were intended to be of **DateTime** or **Date** type, were initially formatted as objects. These columns were subsequently converted to the appropriate **DateTime** format

• String values were cleaned by trimming spaces in all text columns and capitalizing the **State** column to standardize the data, ensuring consistency and improving the quality of reporting.

• The Gender column lacked consistent population of values, likely due to the free-text input feature in the application, allowing users to manually enter their response. The values in the Gender column were as follows:

Gender	**Count**
female	64240
male	25829
transgender	1772
prefer_not_to_say	1350
non_binary	473
unknown	196
not_listed	180
Non-Binary	34
not_specified	28
My gender isn't listed	5
Prefer not to say	1

To ensure consistency and improve data accuracy for analysis, I aggregated and standardized similar values. The cleaned **Gender** column now reflects the following values:

Gender	**Count**
female	64240
male	25829
unknown	6301
transgender	1772
Prefer_not_to_say	1351
Non_binary	507

- There are 7,612 users categorized under the Older Generation (1900-1960). Given the nature of our data, this volume appears unusually high and may indicate redundant or inaccurate records. They might be fake records

GENERATION	USER_COUNT
1900-1960 (Older Generation)	7614
1961-1980 (Gen X)	28849
1981-1996 (Millennials)	33758
1997-2012 (Gen Z)	26077
2013+ (Gen Alpha)	3702

Products:

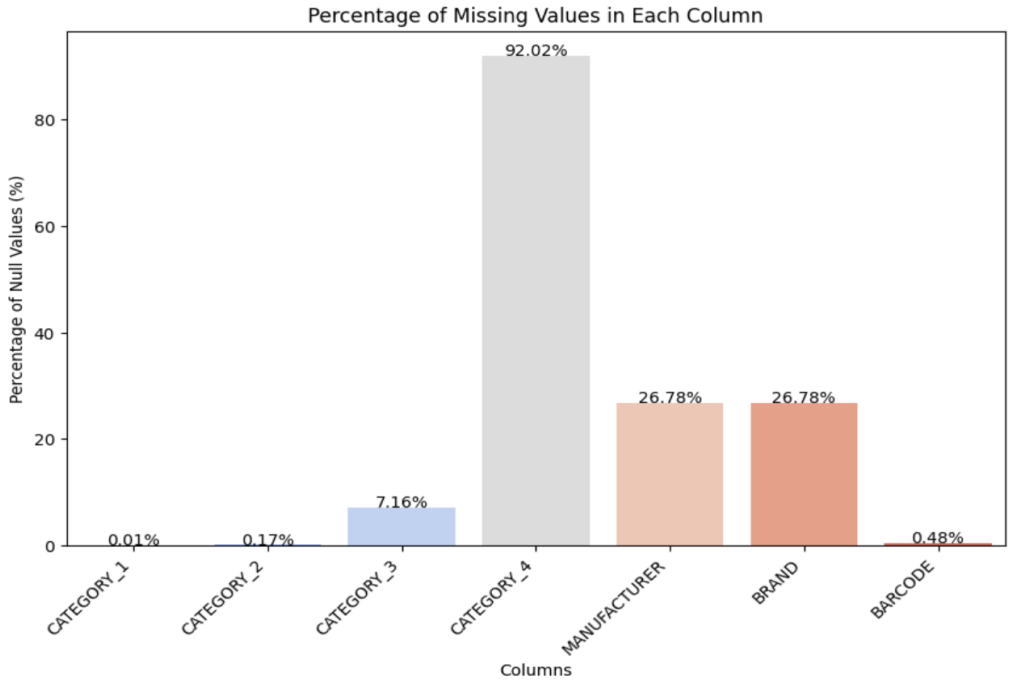
The products table gives us the details of each product, their categories, their brand, manufacturer and barcode. The table also contains hierarchical product categories which might help with flexible filtering

and analysis of products at different levels of detail. Each product might have one or many manufacturers and their associated barcodes.

Data Quality Issues:

• The **Barcode** column, with a .0 appended to each value, was formatted as a float, resulting in **datatype inconsistencies**. To preserve the data and ensure consistency, I first converted the column to an integer and then transformed it into a string format.

• A significant number of **null or missing values** were identified across various columns. Out of the 845,552 records in the table, the following columns contained null values:



Category_4 has almost all of the rows as null, this also needs to be addressed for future purposes for unique product identification.

Rather than deleting these null values, which could result in the loss of valuable data regarding users and their transactions, the nulls were normalized to ensure better understanding and facilitate more accurate analysis.

• Approximately **215 duplicate records** were identified. These duplicates were resolved by retaining the first occurrence of each record and removing the subsequent duplicate entries.

• The **Manufacturer** column contains a significant number of **placeholder values**, with approximately 86,902 entries displaying "PlaceHolder Manufacturer." Similarly, the **Barcode** column includes 17,025 entries marked as "Brand not known." These values indicate that a substantial portion of the data is unknown, which may lead to inaccurate or skewed analysis results.

Transactions:

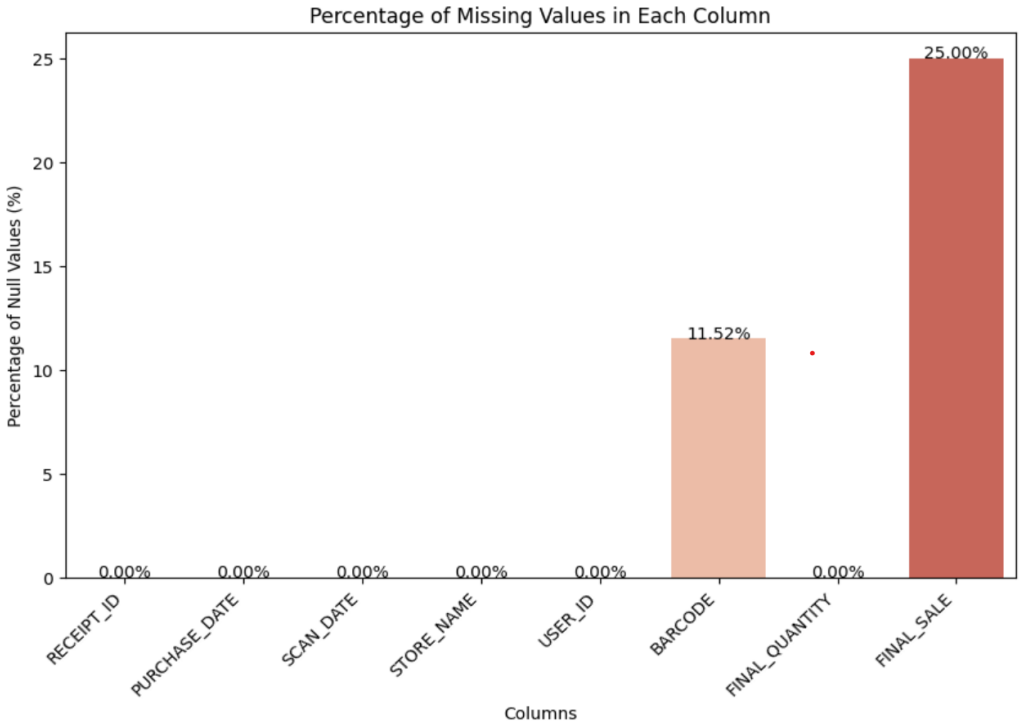
The given transactions table has lifecycle of a receipt such as transaction dates, who scanned the receipt (userid). It also has the items (barcode) associated within the receipt and the total items purchased and money spent on them.

Data Quality Issues:

• Approximately **171 duplicate records** were identified. These duplicates were resolved by retaining the first occurrence of each record and removing the subsequent duplicate entries.

• There were **inconsistencies in data types** across several columns. The **Scan Date** and **Purchase Date** columns, which were intended to be in **DateTime** or **Date** format, were initially set as objects. These were converted to **DateTime** fields to facilitate more accurate analysis of user data. Additionally, the **Barcode** column was converted to a string format, and the **Final Quantity** column, which contained decimal values, was converted to a **float** data type for consistency.

• A significant number of **null or missing values** were found across various columns. Out of 50,000 records in the table, the following is the null distribution.



Fields Challenging to Understand:

• **The Final Quantity values in the Transactions** table exhibit inconsistent behavior. The Final Quantity field contains float/decimal values, suggesting that it may not represent the number of items purchased. It is likely that these values correspond to the weight of the items purchased, rather than item count.

• **The Final Sale values in the Transactions** table contain zero values, despite having a valid **Final Quantity**. This indicates that the customer has purchased goods but did not make a payment. This behavior is unexpected and could be due to factors such as discounts applied to the item or issues with the scanning process, where the item was not properly recorded on the receipt.

• The **Barcode** values in the **Products** table are not unique, as the same barcode is assigned to products from two different manufacturers within the same category. This suggests that the same product may be produced by multiple manufacturers. This issue could stem from data recording errors, or a lack of unique product IDs associated with these barcodes.

Discrepancy due to BARCODE column:

From the given **ER Diagram** we see that ProductsTable is connected to Transactions table through **one-to-many** relationship. But after performing our Exploratory Data Analysis, we found out that BARCODES in Products table for not unique which can disrupt the relationship between these tables and lead to **many-to-many** relationship. The problem with this relationship is that it can lead to exploding joins which might effect accuracy in our analysis.

This problem can be resolved by

- **Establishing a Unique Identifier:** Introduce a new column by concatenating **BARCODE** and **Manufacturer** to create a unique identifier or add a **ProductID** as a distinct key for each product.
- **Enhancing Table Relationships:** Link this unique identifier directly to the **Transactions Table**
- **Implementing a Mapping Table (if needed):** If the **Transactions Table** lacks a **ProductID** column, create a **Mapping Table** containing only **BARCODE** and **ProductID**. This would establish a structured connection:
  Products Table → Mapping Table → Transactions Table, ensuring a properly normalized database structure.
  .