Exercise 1:

Q: Are there any data quality issues present?
Are there any fields that are challenging to understand?

Upon examining the csv files following data quality issues were found in the data.

Products:

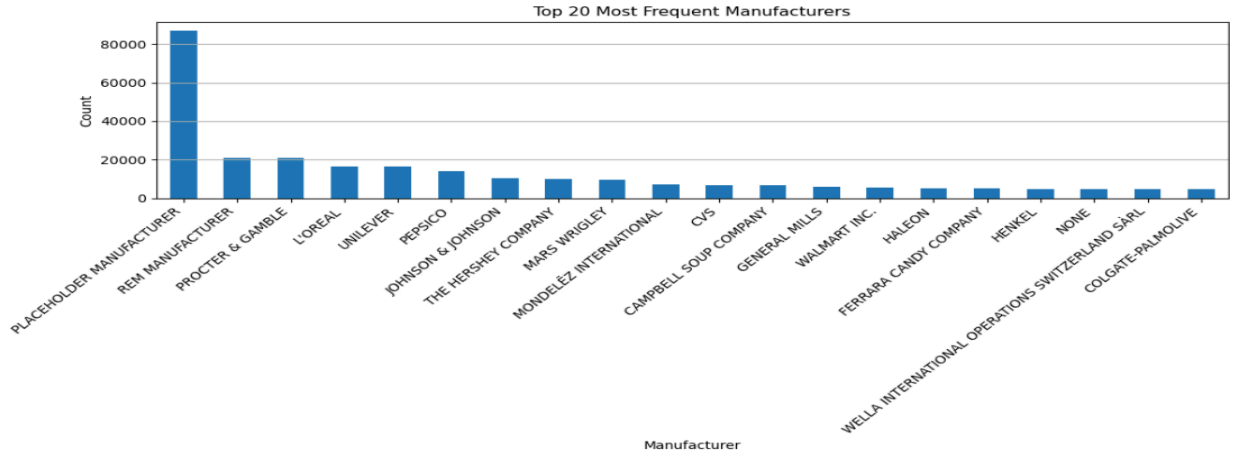
- Since products are part of the dimension tables we need to assure that there are no duplicate records and there are no null values in the primary key of the table which should be BARCODE.
- Since BARCODE was read as float - changed the data type of it to string and appended leading zeros to make all the records uniform 14 digit barcodes.
- After de-duplicating the fully duplicated records and ensuring there are no null values for the barcodes, there were additional barcodes that were duplicated but had different values for other columns. Since there was no deterministic factor to resolve these duplicated barcodes removed these records to ensure data integrity.

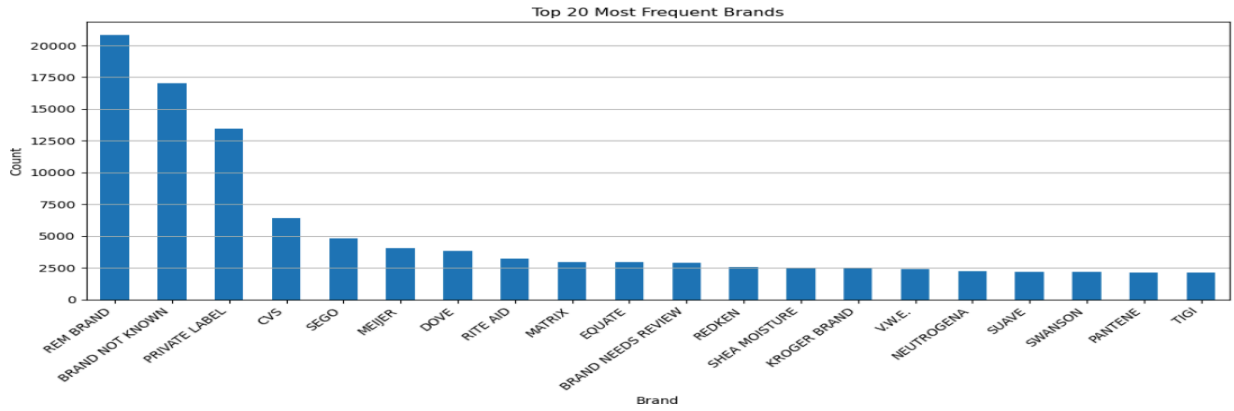
Next checked for missing values in all the columns:

	Missing Count	Missing Percentage
CATEGORY_1	111	0.01
CATEGORY_2	661	0.08
CATEGORY_3	58712	6.98
CATEGORY_4	774085	92.01
MANUFACTURER	226212	26.89
BRAND	226210	26.89
BARCODE	0	0.00

- Category 4 is extremely parse and is not the best value to assess any trends
- Manufacturer and Brand also have ~27% of the data missing

Further, columns MANUFACTURER and BRAND have ambiguous values like ‘Placeholder Manufacturer’, ‘REM Manufacturer’ and ‘REM Brand’ and ‘Brand Not Known’ respectively for a majority of barcodes





Users:

- Fixed data type for columns BIRTH_DATE (date) and CREATED_DATE(datetime) to maintain data quality.

- Next checked for missing values in all the columns:

	Missing Count	Missing Percentage
ID	0	0.00
CREATED_DATE	0	0.00
BIRTH_DATE	3675	3.68
STATE	4812	4.81
LANGUAGE	30508	30.51
GENDER	5892	5.89

Transactions:

- Removed fully duplicated rows resulting in dropping 171 records.
- Fixed format of BARCODE column making it similar to Products dimension table (14 digit string with leading zeros)

- Next checked for missing values in all the columns:

	Missing Count	Missing Percentage
RECEIPT_ID	0	0.00
PURCHASE_DATE	0	0.00
SCAN_DATE	0	0.00
STORE_NAME	0	0.00
USER_ID	0	0.00
BARCODE	5735	11.51
FINAL_QUANTITY	0	0.00
FINAL_SALE	0	0.00

- Column FINAL_QUANTITY has string value ‘zero’ instead of 0.0. Changed data type of the column after fixing the values to float.
- Fixed data type for columns PURCHASE_DATE (date) and SCAN_DATE(datetime) to maintain data quality.
- 8 records had barcode values as -1 - removed these records to maintain only valid records.

- Since the composite key of the transactions table should be a combination of 'RECEIPT_ID', 'PURCHASE_DATE','SCAN_DATE','STORE_NAME','USER_ID' and 'BARCODE', there were multiple records for each of this key.
- To deduplicate I leveraged grouping by each of the above columns and looking at the max value for FINAL_QUANTITY and FINAL_SALE to maintain data quality.

The transactions fact table has very poor matches with both the users and products dimensions with only 0.52% user ids and 59% barcodes matching showcasing poor data quality for user and product dimensions.

Exercise 2:

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

BRAND	TOTAL SALES ($)
CVS	72.0
DOVE	30.91
TRIDENT	23.36
COORS LIGHT	17.48
TRESEMME	14.58

Who are Fetch's power users?

Assumption: Only considering transactions fact to find out power users due to poor data quality of the users dimension.

USER_IDS	RECEIPTS_SCANNED
62925c1be942f00613f7365e	10
64063c8880552327897186a5	9
6327a07aca87b39d76e03864	7
609af341659cf474018831fb	7

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

Ans: TOSTITOS is the leading brand in Dips and Salsa Category when compared with Final Sales and No of Receipts scanned.

Comparison	Top Brand	Value
By Final Sales	TOSTITOS	$181.3
By No of Receipts Scanned	TOSTITOS	36

Exercise 3:

Subject: Summary of Data Review – Quality Issues & Key Insights

Hi Team,

I've completed a detailed review of the product, user, and transaction datasets and wanted to share a quick summary of findings that may impact our analysis and next steps.

Key Data Quality Issues:

- **Products:** ~27% of the records are missing manufacturer and brand details. Many entries have ambiguous values like "REM Brand" or "Placeholder Manufacturer." Additionally, several barcodes are duplicated but point to conflicting product details, we removed these to preserve integrity.

- **Users:** A majority of user IDs in the transaction data don't match the user table; only **0.52%** of transaction user IDs link to valid users. This limits our ability to segment or analyze by user demographics.

- **Transactions:** The data had formatting issues (e.g., barcode inconsistencies, "zero" as a string instead of a number) and duplicate records. These were cleaned. However, only **59%** of barcodes in transactions matched with product records, another flag for dimension alignment issues.

Interesting Trend: Even with these issues, one clear trend emerged: **TOSTITOS is the dominant brand in the Dips & Salsa category**, leading both in sales and number of receipts scanned. This suggests strong customer engagement with this brand in the category.

Request: To move forward, we need help with the following:

- Clarification on how barcodes and user IDs are generated and matched across systems.

- Any available documentation on how ambiguous brand and manufacturer values (e.g., "REM Brand", “Placeholder Manufacturer”) are assigned.

- Guidance on whether we should continue using the existing user and product tables for future analysis, or if cleaner versions are available elsewhere.

Happy to discuss further or walk through these findings live if helpful.

Best,
Ibrahim Ahmed