

Fetch Data Review

Data Quality Issues:

1. Products Data Set Key Issues:

- **CATEGORY_4** missing 92% value, making it largely unusable.
- **MANUFACTURER** and **BRAND** fields have substantial **missing** values (**26.78%**), which could affect product-level analysis.

Products Information:

	Data Type	Missing Values	Non-Null Count	% Missing
CATEGORY_1	object	111	845441	0.01
CATEGORY_2	object	1424	844128	0.17
CATEGORY_3	object	60566	784986	7.16
CATEGORY_4	object	778093	67459	92.02
MANUFACTURER	object	226474	619078	26.78
BRAND	object	226472	619080	26.78
BARCODE	float64	4025	841527	0.48

2. Transactions Data Set Key Issues:

- **BARCODE** missing on **11.52% of transactions**, Will create linkage issues between transactions and products table

Transaction Information:

	Data Type	Missing Values	Non-Null Count	% Missing
RECEIPT_ID	object	0	50000	0.00
PURCHASE_DATE	object	0	50000	0.00
SCAN_DATE	object	0	50000	0.00
STORE_NAME	object	0	50000	0.00
USER_ID	object	0	50000	0.00
BARCODE	float64	5762	44238	11.52
FINAL_QUANTITY	object	0	50000	0.00
FINAL_SALE	object	0	50000	0.00

- **FINAL_SALE** field has a **Blank Space** for **25% of Transactions**, this will **highly impact the revenue calculations** and is a major concern.

transactions_outliers

	RECEIPT_ID	PURCHASE_DATE	SCAN_DATE	STORE_NAME	USER_ID	BARCODE	FINAL_QUANTITY	FINAL_SALE
count	50000	50000	50000	50000	50000	50000	50000	50000
unique	24440	89	24440	954	17694	11028	87	1435
top	bedac253-2256-461b-96af-267748e6cecf	2024-06-15	2024-09-08 20:00:42.348 Z	WALMART	64e62de5ca929250373e6cf5	nan	1.00	
freq	12	774	12	21326	22	5762	35698	12500

3. Transactions Data Set Key Issues:

- **LANGUAGE** missing **30.51%** of entries, with a potential to lose key insights.

User Information:

	Data Type	Missing Values	Non-Null Count	% Missing
ID	object	0	100000	0.00
CREATED_DATE	object	0	100000	0.00
BIRTH_DATE	object	3675	96325	3.68
STATE	object	4812	95188	4.81
LANGUAGE	object	30508	69492	30.51
GENDER	object	5892	94108	5.89

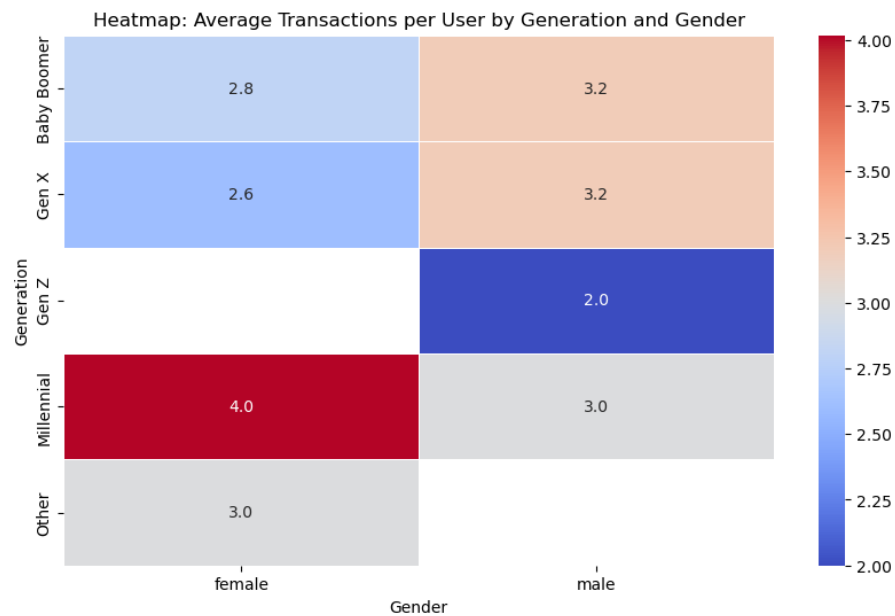
- **1.3%** of the total available **birthdates** seem to be on 1970-01-01 **potential placeholder**
- **BARCODE** misses some transactions, will create linkage issues between products and transactions table.

users_outliers

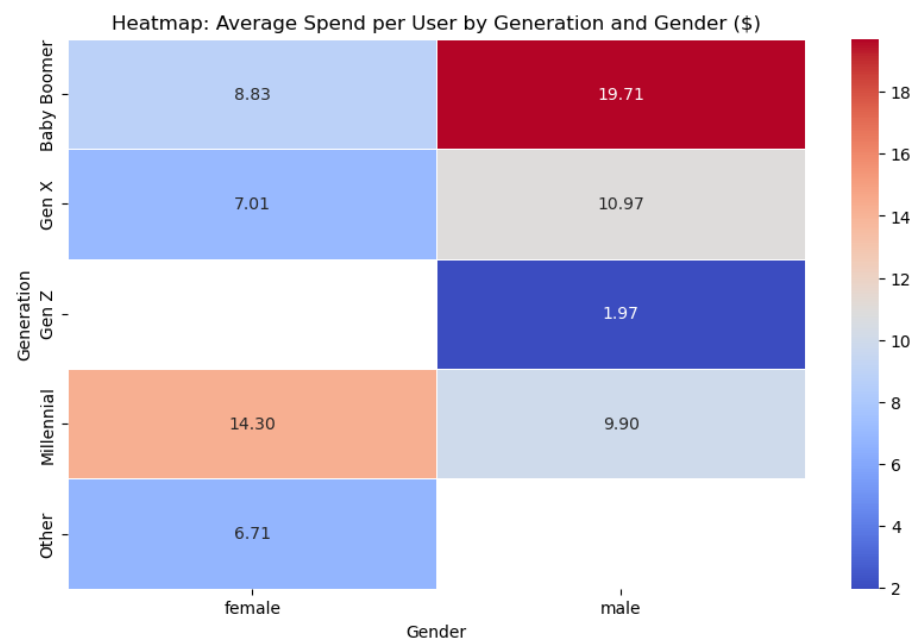
	ID	CREATED_DATE	BIRTH_DATE	STATE	LANGUAGE	GENDER
count	100000	100000	96325	95188	69492	94108
unique	100000	99942	54721	52	2	11
top	5ef3b4f17053ab141787697d	2023-01-12 18:30:15.000 Z	1970-01-01 00:00:00.000 Z	TX	en	female
freq	1	2	1272	9028	63403	64240

Additional Key Insights:

- Millennial Females had the highest number of Average Transactions at 4 followed by GenX & Baby Boomer Males with 3.2



- Whereas the Average Spend per user was at \$19.71 with Baby Boomers Males (61 - 79 Year Olds) 37% higher than the Second highest Millennial Females (29 - 44 Years old) at \$14.30



- Below is a detailed representation of Generation and Product Category Transactions and Spend:
 - Baby Boomer (Males) Health and Wellness has the highest cost per transaction
 - Gen X (Males) have highest transaction rate in Product Categories like Alcohols, Accessories, Beverages

