Zentra Electronics Dataset Summary, Issues, & Caveats

Dataset Summary

Table Summary

Column	Column Name	Description	Data Type
1	USER_ID	User/Customer ID	String
2	ORDER_ID	Order ID	String
3	PURCHASE_TS	Date of order purchase in mm/dd/yy format	Date
4	PURCHASE_TS_CLEANED	Cleaned version of date of order purchase in mm/dd/yy format	Date
5	PURCHASE_MONTH	Date column representing the start of the month for PURCHASE_TS_CLEANED	Date
6	PURCHASE_YEAR	Date column representing the year for PURCHASE_TS_CLEANED	Int
7	SHIP_TS	Date of order shipment in mm/dd/yy format	Date
8	DELIVERY_TS	Date of order delivery in mm/dd/yy format	Date
9	REFUND_TS	Date of order refund in mm/dd/yy format	Date
10	REFUND_TS_CLEANED	Cleaned version of date of order refund in mm/dd/yy format	Date
11	REFUNDED	Binary field for refund or not refund	Boolean
12	PRODUCT_NAME	Name of the product	String

13	PRODUCT_NAME_CLEANED	Cleaned version of name of the product	String
14	PRODUCT_ID	Product ID	String
15	USD_PRICE	Price of the order (in USD currency)	Decimal
16	LOCAL_PRICE	Price of the order (in local currency)	Decimal
17	CURRENCY	Type of currency used in order	String
18	PURCHASE_PLATFORM	Platform used to make the order	String
19	MARKETING_CHANNEL	Type of marketing that led to this order	String
20	ACCOUNT_CREATION_METHOD	Type of device the customer used to create their account	String
21	COUNTRY_CODE	Country in which the order was placed	String
22	LOYALTY_PROGRAM	Binary field to represent if the customer is a loyalty member or not	Boolean
23	CREATED_ON	Date of account creation in mm/dd/yy format	Date
24	REGION	Region in which the order was placed	String

Table Breakdown

Category	Details
Table Summary	**Number of rows**: 108, 127 **Number of columns:** Original - 17; Final - 24
Dataset Description	Order data for Zentra Electronics users from 2019-2022
Grain	Each row represents a distinct order transaction per user
Measures	USD_PRICE, LOCAL_PRICE
Dimensions	USER_ID, ORDER_ID, PURCHASE_TS, PURCHASE_MONTH, PURCHASE_YEAR, SHIP_TS, DELIVER_TS, REFUND_TS, PRODUCT_NAME, PRODUCT_ID, CURRENCY, PURCHASE_PLATFORM, MARKETING_CHANNEL, ACCOUNT_CREATION_METHOD, COUNTRY_CODE, LOYALTY_PROGRAM, CREATED_ON, REGION

Data Issues

Column Issues

Issue ID	Column Name	Issue Description	Magnitude	Resolved?	Resolution Notes
1	ORDER_ID	There are rows where all entries in the columns are the same, but order_id is 1 digit different	169 rows (<1%)	N	No changes were made because we have no source of truth, but made a note for data engineering teams
2	PURCHASE_TS	NULL Values	3 rows (<1%)	N	No changes were made because we have no source of truth, but made a note for data engineering teams
3	PURCHASE_TS	Few entries that have a minute timestamp	27 rows (<1%)	Y	Used Date function to change all cells to m/dd/yy format
4	PURCHASE_TS	Nonsensical (1 record shown as /N)	1 row (<1%)	N	No changes were made because we have no source of truth, but made a note for data engineering teams
5	PRODUCT_NAME	Has some related fields that are not consolidated	197 rows (<1%)	Y	Used find and replace to consolidate 27in”” to 27in
6	PRODUCT_NAME	Inconsistent capitalization	27 rows (<1%)	Y	Used Proper() to capitalize all values

		between product names			
7	USD_PRICE	Nonsensical Values of ($0)	158 rows (<1%)	N	No changes were made because we have no source of truth, but made a note for data engineering teams
8	LOCAL_PRICE	Nonsensical Values of ($0)	158 rows (<1%)	N	No changes were made because we have no source of truth, but made a note for data engineering teams
9	CURRENCY	Missing Values	54 rows (<1%)	N	No changes were made because we have no source of truth, but made a note for data engineering teams
10	MARKETING_CHANNEL	NULL and “Unknown” Values	1387 rows (1.3%)	N	No changes were made because we have no source of truth, but made a note for data engineering teams
11	Account_Created_ON	NULL and “Unknown” Values	1387 rows (1.3%)	N	No changes were made because we have no source of truth, but made a note for data engineering teams
12	COUNTRY_CODE	Missing values	140 rows (<1%))	N	No changes were made because we have no source of truth, but made a note for data engineering teams

13	COUNTRY_CODE	Has some related fields that are not consolidated (NA & North America) in the country_lookup_raw table	2 rows (<1%)	Y	Used find and replace to consolidate North America to NA
14	COUNTRY_CODE	Two rows for the US, one was nonsensical in the country_lookup_raw table	2 rows (<1%)	Y	Nonsensical row was removed
15	COUNTRY_CODE	BJ and BM did not have a region listed in the country_lookup_raw table	2 rows (<1%)	Y	EMEA was added to BJ and LATAM was added to BM using the corresponding ISO country codes

Contextual Issues

- No refunds were recorded from August 2021 to December 2022, indicating a potential data gap. This issue will be brought to the attention of the data team to recover missing refund data and identify the causes for the missing data to ensure accurate performance monitoring and inventory planning
- Some SHIP_TS dates are earlier than the corresponding PURCHASE_TS dates (15 rows, <1%)
- Some DELIVERY_TS dates are earlier than the corresponding PURCHASE_TS dates (14 rows, <1%)
- Some CREATED_ON dates are later than the corresponding PURCHASE_TS dates (8404 rows, 7.7%)
    - Assuming the website doesn't allow for customers to create an account after making an order, this is a large enough issue for the data team to investigate

Caveats

- Removed blank dates (3 rows) for PURCHASE_MONTH from analysis because of the goal to look at trends over time
- Removed blank dates (3 rows) for PURCHASE_YEAR from analysis because of the goal to look at trends over time
- Refund amounts were calculated by multiplying the number of refunds for the product by the product's AOV.
- Bose Soundsport Headphones were not included in the Refund analysis because this product did not have any refunds from 2019 to 2022