DATA CLEANING NOTE

Orders_data TAB:

- **PURCHASE_TS**

- Inconsistent formatting: Used ‘DATE’ function to make format consistent. Added them in new column ‘PURCHASE_TS_CLEANED’.
- Added separate columns, ‘PURCHASE_MONTH’ and ‘PURCHASE_YEAR’, for detailed segmentation and analysis.

- **SHIP_TS & DELIVERY_TS**

- Added new columns, ‘TIME_TO_SHIP’ and ‘TIME_TO_DELIVERY’ to see the number of days taken to ship and deliver. Calculated the date difference from PURCHASE_TS.
- Removed 15 rows (0.01% of the dataset) where shipment or delivery dates are ahead of purchase date.

- **REFUND_TS**

- 2 columns had future date as in year 2025. Used ‘IF’ and ‘TODAY’ function *[=IF (TODAY () <K2,"", K2)]* and taken rows out with future dates.
- Added new binary column ‘REFUNDED’ to sort out which transaction has refund.

- **PRODUCT_NAME**

- Some of ‘27in 4K gaming monitor’ had typos, and ‘bose soundsport headphones’ are all in low capital letters. Cleaned them and added them to ‘PRODUCT_NAME_CLEANED’ column.

- **USD_PRICE**

- 0-dollar transactions: I left it as it is because it might be coupon or discount applied. Total 158 rows (0.1% of the dataset) – Those data have blank cells in CURRENCY column, left them as they are.
- Blank values: I removed cells with blank value, there were 33 rows (0.03% of the dataset) without any value.

- **MARKETING_CHANNEL & ACCOUNT_CREATION_METHOD**

- 1,386 rows (1.3% of the dataset) have missing values. Filled them with 'unknown' to maintain consistency with existing 'unknown' values.

- **COUNTRY_CODE**

- 140 rows (0.1% of the dataset) have missing values. Filled them with 'unknown'.

- **REGION**

- Used ‘V-LOOKUP’ function to bring region to the same sheet. *[=IFERROR (VLOOKUP (W2, country_lookup_raw! A1: B193,2, FALSE), "Unknown")].*