Data Cleaning Documentation

Orders Data Table

ORDER_ID	Identified and filtered out 15K entries with duplicate order IDs.
PURCHASE_TS + PURCHASE_MONTH + PURCHASE_YEAR	Filtered out blank and null entries. (Less than 5 entries) Formatted all dates to a uniform date format. Formatted dates into new purchase month and year columns for use in pivot table breakdown in analysis portion.
PRODUCT_NAME	Consolidated different variations of the same product into a single name. Standardized capitalization among all product names for uniformity.
USD_PRICE	Identified 159 entries with missing prices. Considering the small percentage (less than 0.11%) these missing price entries represented, they were omitted from the analysis.
MARKETING_CHANNEL	Identified 1.4K missing entries and imputed them as “unknown” marketing channel entries.
ACCOUNT_CREATION	Identified 1.3K missing entries and imputed them as “unknown” account creation methods, similarly to the marketing channel column.
REFUND_TS	Identified and filtered out nonsensical values. i.e. Entries with refund dates before their purchase date.
+ DAYS_TO_DELIVER	Added days to delivery column by calculating the difference between purchase date and delivery date columns.

Country Lookup Table

COUNTRY_CODE	Identified and filtered out nonsensical country code values.
REGION	Imputed the correct value for countries missing their corresponding region. Added the region identifier to orders table by joining data with VLOOKUP.

Columns with a (+) represent data added to and calculated from the original data.