

Data Transformation Summary:

◆ `olist_order_items_dataset`

- **Rows:** 112,650 | **Columns:** 7
- `shipping_limit_date` (datetime) was split into **separate date and time columns**.
- Columns `price` and `freight_value` were converted to **Fixed Decimal Number** to maintain financial data consistency.

◆ `olist_orders_dataset`

- **Rows:** 99,441 | **Columns:** 8
- All datetime columns (e.g., `order_purchase_timestamp`, `order_approved_at`, `order_delivered_carrier_date`, `order_delivered_customer_date`, and `order_estimated_delivery_date`) were split into **separate date and time columns**.
- **Time columns were removed** to retain only the relevant date parts for time-series analysis.

◆ `olist_customers_dataset`

- **Rows:** 99,441 | **Columns:** 5
- `customer_city` values were converted to **Proper Case**.
- **Special characters** were removed from city names.
- `customer_state` values were changed from abbreviations (e.g., "SP") to **full state names** (e.g., "São Paulo").
- **Whitespace trimming** was applied to clean data ends.

◆ `olist_geolocation_dataset`

- **Rows:** 1,000,163 | **Columns:** 5
- First row was used as header (as original headers were generic like "Column1", etc.).
- **Duplicate entries** in `geolocation_zip_code_prefix` were removed.
- `geolocation_city` was cleaned for special characters and formatted to Proper Case.
- `geolocation_state` abbreviations were replaced with **full names**.
- **Trimmed whitespaces** to maintain uniformity.

◆ `olist_order_payments_dataset`

- **Rows:** 103,886 | **Columns:** 5
- `payment_value` was converted to **Fixed Decimal Number** for accurate financial calculations.

◆ `olist_order_reviews_dataset`

- **Rows:** 100,000 | **Columns:** 7
- `review_creation_timestamp` and `review_answer_timestamp` were split into **date and time** using delimiter.
- **Time columns** were removed after splitting.

◆ `olist_products_dataset`

- **Rows:** 32,951 | **Columns:** 9
- `product_category_name` was formatted to **Proper Case**.
- **Merged** with `product_category_name_translation` table (72 rows) to get **English category names**.
- Where translations were missing, **custom logic** (if-then-else) was used to **manually add English equivalents**.
- Category name mismatches (e.g., "Pc_Gamer" to "gaming_pc") were corrected.
- **Unnecessary columns** like `product_length_cm`, `product_width_cm`, and other product dimensions were dropped for simplicity.

◆ `olist_sellers_dataset`

- **Rows:** 3,095 | **Columns:** 4
- Cleaned `seller_city` to remove **special characters** and apply Proper Case.
- Expanded `seller_state` abbreviations into **full names**.
- **Whitespace** removed from all string fields.

✧ Summary

These transformations ensured the dataset was clean, consistent, and analysis-ready — supporting accurate **data modeling**, **DAX measure development**, and **visual storytelling** in Power BI.