Data Dictionary – Amazon Sales Dataset

1. Dataset Overview

- **Dataset Name:** Amazon Sales Data
- **Source:** Amazon Sale Report
- **Purpose:** Used for sales analysis, dashboards, KPIs, and reporting

2. Column Definitions

Column Name	Data Type	Example	Description
index	Integer	0	Sequential row index generated during data loading or processing.
order_id	String	405-8078784-5731545	Unique Amazon order identifier. One order ID may appear multiple times if it contains multiple items.
date	Date (MM-DD-YY)	04-30-22	Order date.
status	String	Shipped - Delivered to Buyer	Raw order status provided by Amazon.
fulfilment	String	Merchant	Indicates whether the order is fulfilled by Amazon or Merchant.
sales_channel	String	Amazon.in	Platform where the order was placed.
ship_service_level	String	Standard	Shipping speed or service level selected by the customer.
style	String	JNE3781	Product style or design identifier.
sku	String	JNE3781-KR-XXXL	Stock Keeping Unit representing a specific product variation.
category	String	kurta	Product category.
size	String	3XL	Size of the product ordered.

Column Name	Data Type	Example	Description
asin	String	B09KXVBD7Z	Amazon Standard Identification Number identifying the product.
courier_status	String	unknown	Shipping or delivery status reported by the courier. Missing values are filled with 'unknown'.
qty	Integer	1	Quantity of units ordered for the given SKU.
amount	Float	406.00	Monetary value of the order line.
ship_city	String	MUMBAI	Destination city for shipment.
ship_state	String	MAHARASHTRA	Destination state for shipment.
ship_postal_code	String / Integer	400081	Postal (ZIP) code of the shipping address.
b2b	Boolean	False	Indicates whether the order is Business-to-Business (True) or Business-to-Consumer (False).
status_clean	String	Delivered	Standardized order status derived from the raw 'status' field for reporting and analysis.

3. Notes / Transformations

- Removed unused columns
- Filled missing values in `courier_status` with 'unknown'
- Replaced null values in `amount` column with the average amount
- Standardized column names (lowercase, spaces replaced with dashes)
- Identified and validated duplicate orders
- Created the `status_clean` column for simplified analysis and reporting

4. Valid Values / Categories

- **Status Clean:** Delivered, Pending, In Transit, Failed, Cancelled
- **Order ID:** Unique per order
- **ASIN:** Unique per product

5. Limitations / Remarks

- Some orders may have multiple items (same `order_id` repeated with different `asin`)
- Exact duplicates should be checked before analysis