Dataset 6 – FMCG Beverage Sales – Realistic Transaction Data

Problem Statement: Analyze daily FMCG beverage sales across outlets, brands, pack types, promotions, weather and events, and build EDA and forecasting features.

Grain (each row represents): One row per outlet × brand/pack × date (daily transactional grain).

Notes: Synthetic but realistic daily transaction dataset with missing values, outliers, rare brand category and duplicates for rich EDA practice.

Data Dictionary:

Column Name	Type	Example	Description
transaction_id	String	TXN_000123	Unique identifier for the sales transaction row.
date	Date	2024-07-15	Calendar date for the sales transaction (year 2024).
region	Category	West	Macro region: North, South, East, West.
city	Category	Mumbai	City where the outlet is located.
outlet_type	Category	Supermarket	Type of outlet: Supermarket, Hypermarket, Convenience Store, Bar/Restaurant, E-commerce.
sales_channel	Category	Off-Trade	Sales channel classification: On-Trade (bars/restaurants) or Off-Trade (retail/e-commerce).
customer_type	Category	Retail Chain	Customer/account type: Retail Chain, Independent Store,

			HoReCa, Modern Trade.
brand	Category	Heineken	Brand sold in the transaction (Heineken, Kingfisher, Budweiser, Local_Brand_A, Local_Brand_B, Imported_Special).
pack_size_ml	Numeric (integer)	650	Pack size in milliliters (e.g. 330, 500, 650, 1000).
pack_type	Category	Bottle	Packaging type: Can, Bottle, Keg, Draft.
promo_flag	Numeric (0/1)	1	1 if a promotion is active for this transaction, 0 otherwise.
promo_type	Category	Price_Off	Type of promotion: None, Price_Off, Bundle, Display, Digital_Coupon. Contains missing values for some rows.
discount_pct	Numeric (float)	0.15	Discount percentage applied on list price. Typically 0–0.30, with some outliers at 0.60; contains missing values.
competitor_price_index	Numeric (float)	1.05	Relative competitor price index (~0.8–1.2); >1 means competitors are more expensive. Contains missing

			values.
units_sold	Numeric (integer)	55	Number of units sold in the transaction. Includes some extreme outliers for teaching EDA.
base_price_per_unit	Numeric (float)	165.50	Base list price per unit before discount, driven by brand, pack size and pack type.
gross_sales_value	Numeric (float)	9102.50	Gross sales value before discounts: units_sold × base_price_per_unit.
net_sales_value	Numeric (float)	7737.13	Net sales value after discounts: gross_sales_value × (1 – discount_pct).
stock_at_start	Numeric (integer)	120	Stock units available at the start of the day at this outlet for this SKU/brand.
stock_at_end	Numeric (integer)	40	Stock units remaining at the end of the day.
oos_flag	Numeric (0/1)	0	Out-of-stock flag: 1 if stock_at_end is zero, else 0.
weather	Category	Sunny	Weather condition on that day: Sunny, Cloudy, Rainy, Humid, Cool. Contains missing values.

temp_c	Numeric (float)	29.5	Approximate temperature in °C for that region/city on that day. Contains missing values.
event	Category	Festival	Local event type impacting demand: None, Festival, Sports_Event, Concert, Long_Weekend.