Heineken Analytics Datasets – Data Dictionary & Descriptions

Dataset 1 – Daily Sales Forecasting for 2 Heineken SKUs

Problem Statement: Forecast daily sales (units and value) for two key Heineken SKUs across regions and channels, using promo, events, competitor index and temperature as drivers.

Grain (each row represents): One row per day × region × channel × SKU.

Notes: Dataset includes missing values, outliers in units_sold and net_sales_value, and duplicate rows to support EDA and data cleaning exercises.

Data Dictionary:

Column Name	Type	Example	Description
date	Date	2024-07-15	Calendar date of the transaction (daily data for 2024).
region	Category	West	Macro region: North, South, East, West.
city	Category	Mumbai	City where sales happen, mapped to region.
channel	Category	Off-Trade	Sales channel: On-Trade (bars/restaurants) vs Off-Trade (retail/e-com).
sku	Category	Heineken_330_Can	Product identifier: Heineken_330_Can or Heineken_650_Bottle.
list_price_per_unit	Numeric (float)	160.5	Base list price per unit before discount, with small random variation.
promo_flag	Numeric (0/1)	1	1 if a promotion is active for that SKU/region/channel/day, else 0.
promo_discount_pct	Numeric (float)	0.18	Discount percentage applied if promotion is active. Contains missing

			values for some rows.
competitor_price_index	Numeric (float)	1.03	Relative competitor price index (~0.85–1.15). Values >1 mean competitors are more expensive. Contains missing values.
temp_c	Numeric (float)	29.2	Approximate temperature (°C), with seasonal pattern and noise. Contains missing values.
event	Category	Festival	Event type for that day: None, Festival, Sports_Event, Long_Weekend.
units_sold	Numeric (integer)	120	Units sold for that combination of date, region, channel, and SKU. Includes some injected outliers (very high days).
net_sales_value	Numeric (float)	18250.75	Realized net sales value = units_sold × list_price_per_unit × (1 – promo_discount_pct).