DATA DICTIONARY

Myntra Analytics Dataset

This document describes all columns present in the dataset used for this project.

Source File

`data/processed/myntra_cleaned_data.xlsx`

Column Reference

Column Name	Data Type	Description	Example Value
`product_id`	String / Integer	Unique identifier for each product listing	`10001`
`product_name`	String	Full product name as listed on Myntra	`"Roadster Men Slim Fit Jeans"`
`brand`	String	Brand of the product	`"H&M"`, `"Roadster"`
`category`	String	Top-level product category	`"Men's Clothing"`
`sub_category`	String	More specific product type	`"Jeans"`, `"T-Shirts"`
`original_price`	Float	Listed MRP before any discount (₹)	`1499.00`
`discounted_price`	Float	Final selling price after discount (₹)	`899.00`
`discount_pct`	Float	Percentage discount applied (derived column)	`40.0`
`rating`	Float	Average customer rating on a 1–5 scale	`4.2`
`rating_count`	Integer	Number of customers who submitted a rating	`1523`
`price_band`	String	Derived: price tier bucket	`"Budget"`, `"Mid-range"`, `"Premium"`
`rating_tier`	String	Derived: rating category	`"High (4+)"`, `"Mid (3-4)"`, `"Low (<3)"`

Note: Derived columns (`discount_pct`, `price_band`, `rating_tier`) were created during the Excel cleaning phase and are not present in the raw dataset.

Data Quality Notes

- **Missing values:** Rating and rating_count were missing for approximately 23% of products — these were retained but filtered out in rating-specific visuals.
- **Duplicates:** 83164 duplicate product IDs were found and removed during cleaning.
- **Price anomalies:** 376 records had a discounted price higher than original price — these were reviewed and corrected/removed.
- **Category standardisation:** Category names were normalised to consistent casing and spelling.

Price Band Definitions

Band	Price Range (₹)
Budget	< ₹500
Mid-range	₹500 – ₹2,000
Premium	> ₹2,000

Last updated: 2025