Aleksandar Stojov

Project Deliverable #1

This dataset contains detailed information on over 42,000 Amazon electronics products, including sales figures, customer ratings, pricing trends, and sub-category classifications. It is provided in two versions: cleaned file - fully preprocessed and ready for analysis, visualization, and machine learning projects and uncleaned file - Raw scraped data for practicing data cleaning, preprocessing, and feature engineering.

With 17 enriched features, this dataset is well-suited for applications in data science, machine learning, recommendation systems, and business analytics.

Key columns:

product\_title – Complete name/title of the product

product\_rating – Average customer rating (numeric) out of 5

total\_reviews – Total number of customer reviews

purchased\_last\_month – Units purchased in the last month

discounted\_price – Current price after discount

original\_price – Original listed price before discount

discount\_percentage – Percentage discount applied to the product

is\_best\_seller – Indicates if the product is tagged as a Best Seller

is\_sponsored – Whether the product is a Sponsored item or Organic

has\_coupon – Special discounted coupons availability (True/False)

buy\_box\_availability – BuyBox button availability on amazon search page like add to cart(nan values represent False)

delivery\_date – Estimated delivery date (converted to datetime format)

sustainability\_tags – Eco-friendly and sustainability-related tags

product\_image\_url – Direct image link of the product

product\_page\_url – Official Amazon product page URL

data\_collected\_at – Date when the data was collected

product\_category – Assigned product category based on the title

According to the author, this dataset is useful as it opens a wide range of practical use cases, such as:

* Price Analysis & Trends – Study pricing behavior, discounts, and seasonal sales
* Customer Behavior Analysis – Analyze ratings, reviews, and sales patterns
* Recommendation Systems – Build personalized product recommendation engines
* Market Basket Analysis – Identify related products frequently bought together
* Predictive Modeling – Forecast sales, demand, and discount impact
* NLP Projects – Use product titles for text classification and category prediction
* Data Cleaning Practice – Use the raw file for real-world preprocessing exercises

|  |  |
| --- | --- |
| File format | CSV |
| Structure | Tabular |
| Data types | Numeric and categorical |
| Measurement units | Units sold, ratings, prices |
| Granularity | Individual product listings |
| Dimensions | Over 42000 rows and 17 columns |
| Temporal details | As of 2025 |
| Geographic details | Global listing |

Research questions:

1) Which product categories have the highest sales?

Response variable: Sales

Explanatory variable: Category

Purpose: Determine which categories are most popular or generate the most revenue.

2) Does product price influence sales volume?

Response variable: Sales

Explanatory variable: Price

Purpose: Determine if the price affects sales numbers.

3) Do product ratings affect sales performance?

Response variable: Sales

Explanatory variable: Rating

Purpose: Determine if highly rated products sell better than lower-rated products.