Amazon India Sales Analysis

Soren Epple, Lucas Sadoulet, Tjoet Lakmana

2025-05-01

Introduction

Shopping online has become a huge part of everyday life, and platforms like Amazon make it easy for people to browse, compare, and buy products in just a few clicks. The Amazon India Sales dataset gives us a behind-the-scenes look at how product pricing, discounts, and customer ratings play out across different types of items. With many products and reviews, it's a great dataset to explore what matters to online shoppers.

Here are the main questions we set out to explore:

- How many ratings do products get per customer, and what does that tell us about engagement?

  This helps us understand how actively customers interact with different types of products. A high number of ratings might indicate strong customer engagement or high sales volume, while low counts may suggest limited interaction or visibility.

- How does price vary across different product categories?

  Exploring pricing by category helps us see how products are positioned in the market. Some categories might lean toward luxury pricing, while others depend on affordability and high-volume turnover.

- Is there a connection between discount percentage and customer ratings?

  This question lets us explore whether discounts influence satisfaction. Do big discounts lead to better reviews or do they come at the cost of perceived value or quality?

These questions are important because they help uncover patterns in online shopping behavior. By analyzing engagement, pricing, and satisfaction, we can better understand what makes a product stand out in a crowded marketplace.

Data Provenance

The dataset we used comes from Kaggle, a popular online platform for data science and machine learning. Kaggle hosts thousands of datasets that are free to use for analysis, modeling, and competitions. It is widely used by students, researchers, and professionals to practice and share data projects.

The specific dataset we used is called the "Amazon Sales Dataset," published by Karkavel Rajaj. It includes a wide range of product-related information from Amazon India. the key columns included in the dataset:

- **product_id**: Unique identifier for each product
- **product_name**: Name of the product
- **category**: Product category
- **discounted_price**: Price after discount
- **actual_price**: Original price before discount
- **discount_percentage**: Percent discount applied
- **rating**: Star rating given to the product
- **rating_count**: Number of people who rated the product
- **about_product**: Description or summary of the product
- **user_id**: Unique ID of the user who left a review
- **user_name**: Name of the reviewing user
- **review_id**: ID for the specific review
- **review_title**: Title of the user's review
- **review_content**: Full text of the user's review
- **img_link**: Image URL of the product

- **product_link**: Link to the product listing on Amazon

Data Cleaning and Wrangling

To prepare the data for analysis, we followed a multi-step process to clean, convert, and restructure the data for meaningful insights:

Removed Duplicate Product Entries

We began by removing duplicate entries based on the `product_id` column using `distinct(product_id, .keep_all = TRUE)`. This step ensured that each product was represented only once in our dataset. Duplicates likely existed due to multiple reviews or variations for the same product. Since we were analyzing product-level trends (e.g., pricing, ratings), keeping only the first unique entry per product helped prevent double-counting.

Converted Prices from INR to USD

The dataset included prices in Indian Rupees with formatting like the symbol and commas. We:

- Removed those symbols using `gsub()`
- Converted values into USD using a fixed exchange rate of 1 = \$0.012
- This allowed us to compare prices and discounts using a more familiar currency.

Cleaned and Converted Rating Counts

We cleaned the `rating_count` column by removing commas and converting it to a numeric type. This made it possible to perform calculations like total reviews or average engagement per category.

Selected Relevant Columns

To simplify our dataset and focus on the most useful information, we retained:

- `product_name`, `category`, `discounted_price_USD`, `actual_price_USD`, `rating`, `rating_count`, and `user_id`.

This reduced clutter and made downstream analysis more efficient.

Seperated User IDs

In the raw dataset, some products were associated with multiple `user_id`s stored in a single cell, separated by commas. This structure made it difficult to analyze individual customer interactions. To resolve this, we used the `separate_rows()` function to split these entries so that each `user_id` had its own row.

This step allowed us to more accurately analyze customer engagement by treating each user as a distinct data point, which is especially important for questions involving rating count per customer or overall review behavior.

Final Tidied Data

After completing the cleaning and transformation steps, our final dataset is organized in a tidy format where each row represents a unique interaction between a product and a user. The data includes only the most relevant variables needed to answer our research questions, with all prices converted to USD and formatted consistently.

The columns in the final dataset include:

- **`product_name`** – Name of the product
- **`category`** – Full category path of the product
- **`discounted_price_USD`** – Discounted price in U.S. dollars
- **`actual_price_USD`** – Original price in U.S. dollars
- **`rating`** – Customer rating for the product
- **`rating_count`** – Number of customer ratings
- **`user_id`** – Individual customer ID, one per row

This structure allows us to analyze product pricing, discount trends, rating patterns, and customer engagement at both the product and category levels. By simplifying and standardizing the dataset, we made it easier to create visualizations, perform summaries, and answer the key questions guiding our analysis.

FAIR Principles

The dataset we used for this project aligns with what are known as the FAIR Principles. These principles help guide how data should be handled to make it more usable and shareable across different settings.

Findable

You can find this dataset easily on Kaggle, which is a popular platform for sharing and discovering datasets. However, this specific dataset does not come with detailed metadata or column explanations, so some interpretation and exploration are needed to fully understand its structure. Still, it is clearly labeled and easy to locate on the platform.

Accessible

Getting access to this data is straightforward. There are no paywalls or complicated steps, you just need a free Kaggle account. That means anyone interested in exploring the dataset, whether for school or personal projects, can do so easily.

Interoperable

The data is in CSV format, which is a standard that works with all sorts of tools like R, Python, Excel, and more. This makes it convenient to load into whatever program you're most comfortable using.

Reusable

This dataset is shared in a way that encourages others to use it. Because it's clearly structured and documented, it's great for reuse in class projects, personal analysis, or even more advanced research.

CARE Principles

The CARE Principles are all about making sure data involving Indigenous communities is handled ethically and respectfully. In this case, the dataset doesn't include any cultural or community-specific information, so the CARE guidelines don't really apply here.

Data Visualizations

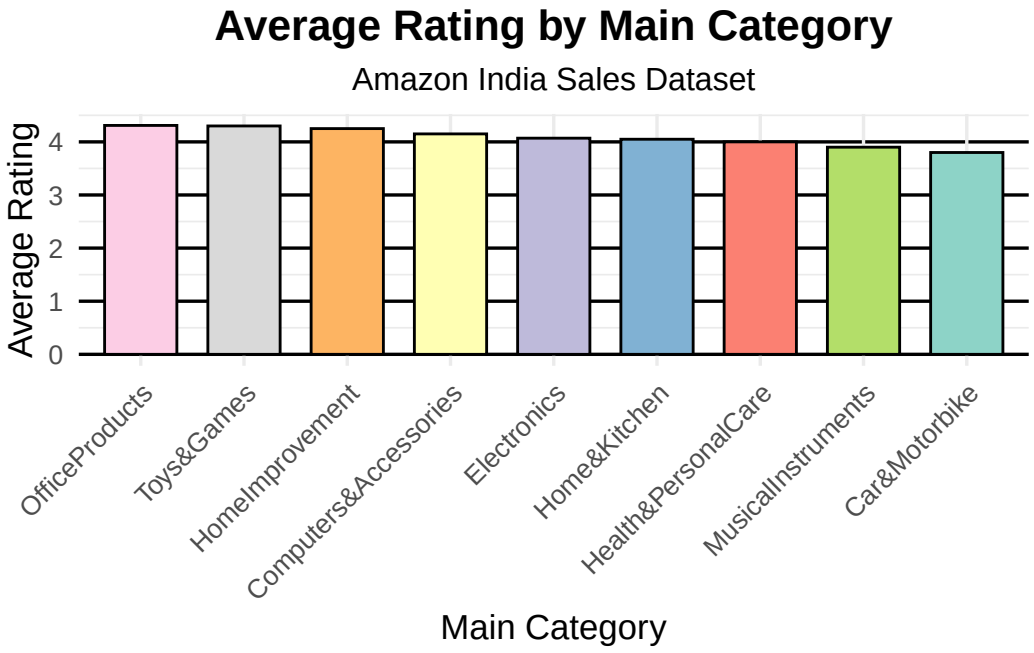
Table 1: Average Rating and Total Reviews by Main Category

main_category	avg_rating	total_reviews
Electronics	4.07	3810
Home&Kitchen	4.05	3513
Computers&Accessories	4.15	2975
OfficeProducts	4.31	248
HomeImprovement	4.25	16
MusicalInstruments	3.90	16

Table 1: Average Rating and Total Reviews by Main Category

main_category	avg_rating	total_reviews
Car&Motorbike	3.80	8
Toys&Games	4.30	8
Health&PersonalCare	4.00	4

Figure 1: Average Rating by Main Category



Report

With the following data visualizations provided above via Table 1 and Figure 1, provides a comprehensive overview of customer satisfaction across various products. While the bar chart displays average rating for each main category, allowing for easy visual comparison, the table supplements this with total review counts, which gives context to the statistical reliability of those ratings.

Key Observations:

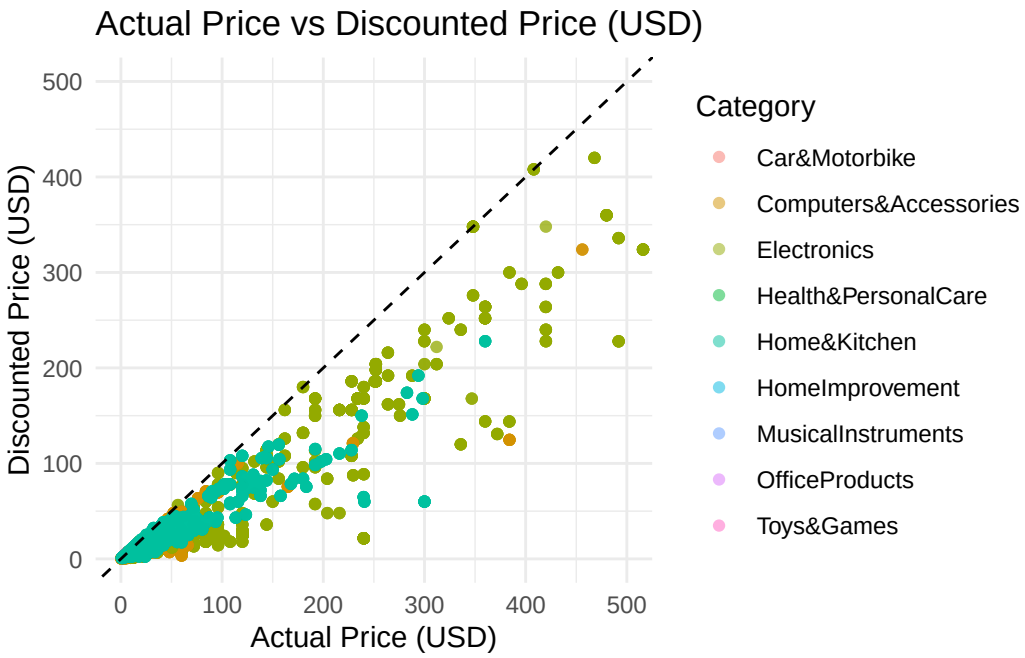
- While *Office Products (4.31)* and *Toy & Games (4.30)* lead in the average rating in customer satisfaction, these categories have relatively low review counts (248 and 8). This may cause future customers to be hesitant as they may be unsure about the reliability of the reviews.
- Categories receiving high volumes in rating count, such as *Electronics (3,810)* and *Home & Kitchen (3,513)*, indicate a larger representative of the customer sentiment and represent a solid satisfaction within high-volume product areas.

- *Car & Motorbike (3.80)* and *Musical Instruments (3.90)* received the lowest ratings across all main categories, though receiving low volumes in review count. Similar to Office Products and Toy & Games, future customers may be hesitant in purchasing; however, a product that received a low rating off the bat may be more likely to cause a chain of unhappy customers.

Final Interpretation:

Categories receiving relatively high average ratings should be interpreted alongside the number of reviews. A product with a high rating and low volume of reviews may be less reliable than a slightly lower rating with a greater volume in reviews. The display of these two visualizations help identify which categories are highly rated, more specifically which ratings are statistically more robust. *Electronics* and *Home & Kitchen* remain at a high volume and consistent with their high ratings.

Figure 2: Actuacl Price vs Discounted Price (USD)



Report

The scatter plot Figure 2 visualizes the relationship between a product's actual price and its discounted price across different categories. Each point represents a product, colored by the main category, and the dashed line (where actual = discounted) acts as a reference for no discount.

Key Observations:

- Most points below the diagonal line, confirming that discounts are widespread.
- Some categories like *Electronics* and *Home & Kitchen* show a wide spread of discount amounts, suggesting more aggressive and varied pricing strategies.

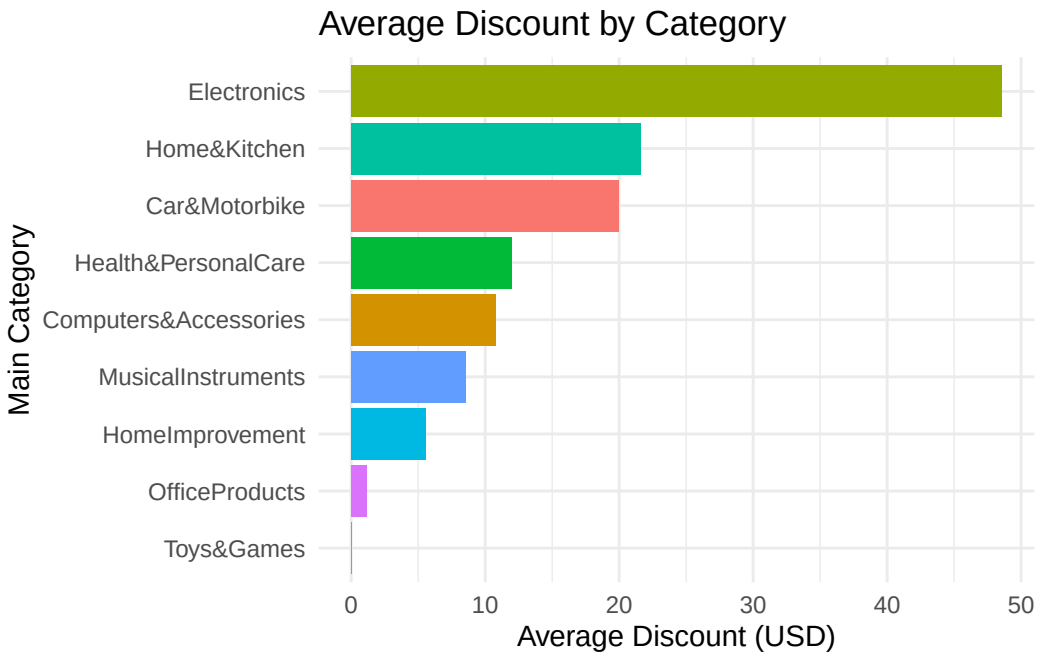
- *Office Products* and *Toys & Games* cluster closer to the diagonal, indicating that these products are less frequently discounted or have smaller discount margins

Final Interpretation:

This plot helps uncover which categories tend to be marketed with more competitive pricing. Categories with wide discount gaps may use pricing as a key strategy to attract buyers, while those closer to full price may rely more on brand value or niche appeal. This aligns with the earlier analysis of customer satisfaction price and perception often go hand-in-hand.

To clearly visualize how discounts vary across categories, we created a bar plot that ranks product categories from those with the highest average discounts to those with the lowest:

Figure 3: Average Discount by Category



Report

Figure 3 presents a horizontal bar chart illustrating the average discount in USD for each main-level product category. By calculating the difference between actual and discounted prices, and then averaging these values per category, the chart allows us to compare how aggressively each category is discounted.

Key Observations:

- Categories at the top of the chart (*Electronics, Home & Kitchen, Car & Motorbikes)* represent those with the largest average discounts, indicating sellers may rely heavily on promotions in these markets to stay competitive

- categories toward the bottom (*Toy & Games, Office Products, Home Improvement)* reflect smaller or less frequent discounts, possibly signaling more stable pricing, stronger brand value, or lower competitive pressure
- Some categories may cluster closely together in average discount, suggesting consistent pricing behavior across similar product types.

Final Interpretation

This visualization highlights how pricing strategies vary across product categories. Categories like *Electronics* and *Home & Kitchen* stand out for offering some of the highest average discounts, and referring to Figure 1 they also maintain high volumes of reviews and consistent average ratings, as seen in the earlier analysis. This correlation suggests that in certain high-volume categories, deep discounts do not undermine customer satisfaction, in fact, they may enhance it by making popular, well-reviewed products more accessible. On the other hand, categories with lower discounts and fewer reviews may appear more premium or niche but lack the broader customer validation. Together, these insights reinforce that discounting, when paired with quality and demand, can still result in strong customer engagement and satisfaction.