Amazon Product Sales & Pricing Analysis

- Aman Sahu

Business Problem

E-commerce platforms struggle to answer:

- What drives product success: price, discount, ratings, or popularity?
- Are high discounts actually helping?
- Are ratings trustworthy?
- Can we predict ideal selling price?

Objective

Analyze Amazon product data to understand:

- Pricing vs discount dynamics
- Impact of ratings & reviews
- Identify category performance
- Build a ML model to predict discounted price

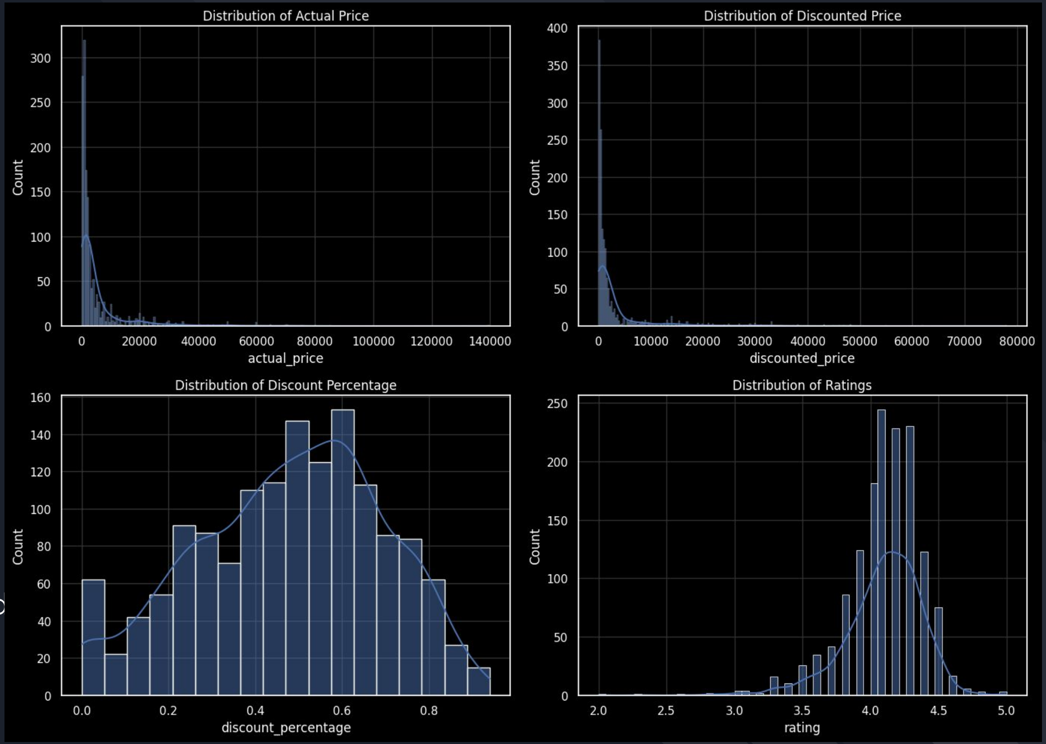
Dataset Overview

Dataset Includes

- Product details: Name, Category
- Pricing: Actual Price, Discounted Price, Discount %
- Customer Feedback: Rating & Rating Count
- Metadata: Reviews, Links
- Shape : [1465 * 16]

Data Cleaning

- Removed symbols, converted to numeric,
- Handled missing values: Filled rating_count using median
- Standardized features
- Created analytical feature: rounded rating



- Actual price and discounted price both are extremely right skewed. The histogram shows a wic spread, but the bulk of the data sits between 30% and 70%.

- The most frequent discounts appear to be in the 50-60% range, suggesting a strategy heavily reliant on deep discounting rather than small markdowns.

- Most products are rated between 4.0 and 4.5, with very few falling below 3.0. This could indicate high product quality, or potentially rating inflation where users rarely give low scores.

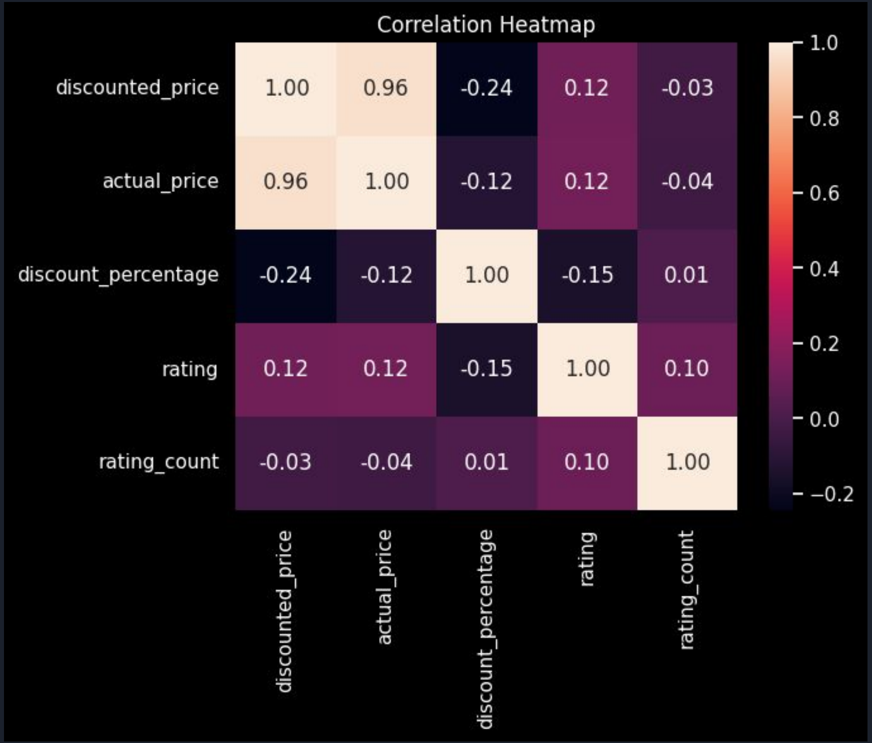
- Higher prices correlate with consistent quality, as products priced over 40,000 rarely drop below a 3.5-star rating.
- In contrast, the budget market is a gamble, where low prices lead to highly volatile customer satisfaction ranging anywhere from 2.0 to 5.0 stars.

Descriptive Statistics

Exploratory Insights

- Market is highly discount-driven
- Median Discount = 50%
- Pricing distribution is right-skewed
- Most products priced between 300–4000
- Ratings centered between 4.0 – 4.5

- Negative correlation (-0.15) between discount_percentage and rating, this suggests that heavily discounted items tend to have slightly lower ratings, perhaps indicating they are clearance stock or perceived as lower quality.
- While the scatter plot shows high-priced items (>$40k$) rarely fail, the overall correlation between actual_price and rating is only 0.12. This confirms that for the vast majority of inventory : the budget tier, paying more does not guarantee a better rating.

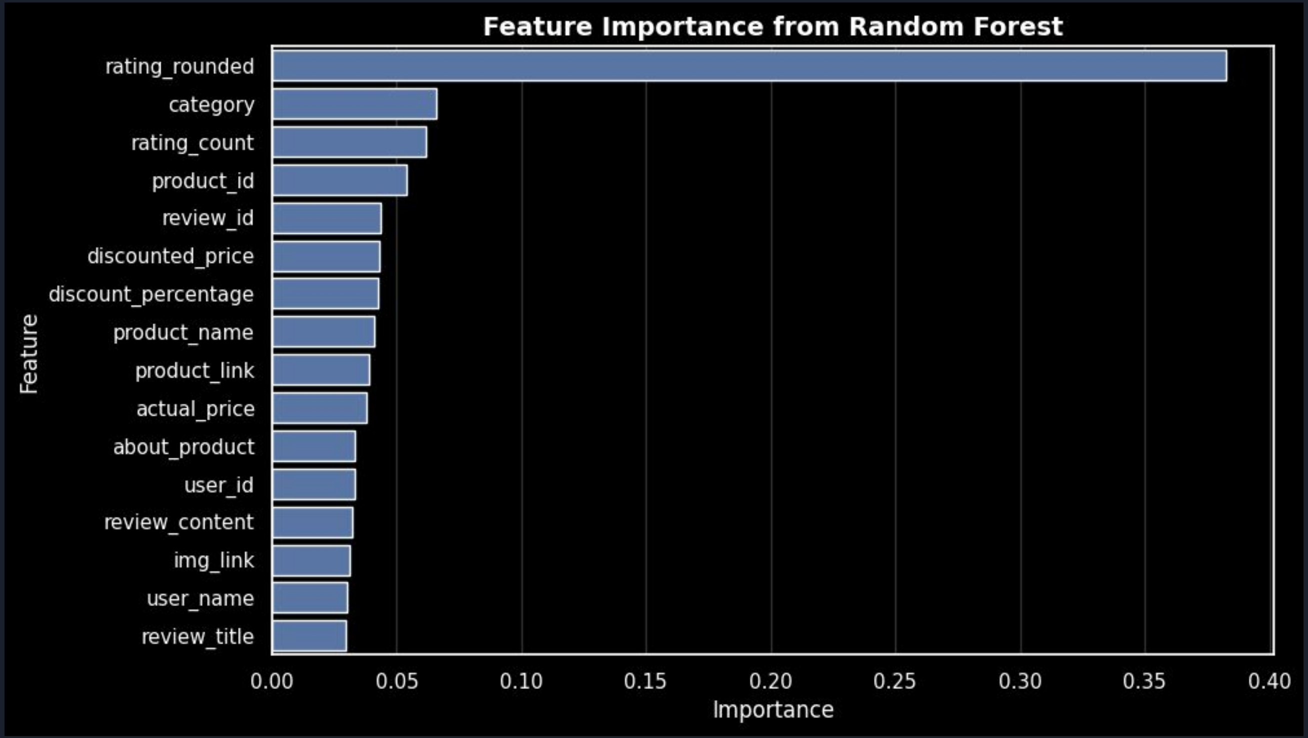


Descriptive Statistics

Ratings & Price Insights

- 95% products rated ≥ 4 stars → **Strong rating inflation**
- Rating alone is not a good performance differentiator
- **Rating Count is more powerful (indicates popularity)**
- Slight negative correlation between discount% & rating → **heavily discounted products are slightly lower rated**
- High-priced premium products rarely fall below 3.5 stars

Popularity + trust matters more than rating score itself.


Feature Importance from Random Forest
rating_rounded
category
rating_count
product_id
review_id
discounted_price
discount_percentage
product_name
product_link
actual_price
about_product
user_id
review_content
img_link
user_name
review_title
Feature
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Importance

Modelling & it’s Performance

Model	RMSE	MAE	R^2
Linear Regression	0.2557	0.1205	0.9377
Decision Tree	0.2439	0.0550	0.9433
Random Forest	0.1897	0.0420	0.9657
Gradient Boosting	0.1743	0.0404	0.9711
XGBoost	**0.1653**	**0.0345**	**0.9740**

Best Model: XGBoost
→ Highest accuracy
→ Lowest error
→ Strong generalization

Model Validation :
Actual vs Predicted plot shows tight alignment
Residuals centered around zero
Model does not underfit / overfit significantly

Conclusion: The model is highly reliable for pricing prediction.

Business Recommendations

- Maintain 50–60% discount strategy, aligns with market expectation
- Focus on increasing rating engagement, not rating value
- Promote popular + trustworthy items (high rating_count)
- Premium segment is stable → can sustain pricing
- Budget products need quality control to reduce volatility

Conclusion

- Amazon market is strongly discount-dependent
- Ratings alone don’t determine success → Engagement matters more
- Pricing behavior varies widely by segment