Sarah Taghadosi
400222019
March 23, 2025

Amazon Sales Analysis

Statistical Report

Abstract

This report presents a statistical analysis of an Amazon product dataset. By applying various statistical tests—including Spearman Correlation, Chi-Square Test, Independent Samples T-test, ANOVA, and the Shapiro-Wilk Test—we explored the relationships between pricing, discounts, product categories, and customer ratings. Visualizations were used to support the findings. Results show complex dynamics in how pricing and categories impact ratings, providing insight into data-driven pricing and marketing strategies.

1. Introduction

This report presents an analysis of Amazon product data to examine whether product pricing, discounts, and categories are associated with customer ratings. Using five statistical hypothesis tests, we aim to understand how these factors relate to customer behavior. The structure of this report follows each research question individually, supported by statistical results and visualizations.

2. Research Questions

1. Does the discounted price significantly impact product ratings?
2. Are product categories and high/low ratings independent?
3. Is there a significant difference in ratings between high- and low-discount products?
4. Do more expensive products receive higher ratings?
5. Does the distribution of rating counts follow a normal distribution?

3. Data

The dataset contains 1,465 products from Amazon, including the following fields: product name, actual price, discounted price, discount percentage, rating, rating count, and product category.
Preprocessing steps included:

- Removing special characters (₹, %, commas)
- Converting price, discount, and rating to numeric
- Dropping missing or invalid entries

4. Methods

Purpose	Test
To assess the relationship between discounted price and rating	Spearman Rank Correlation
To test independence between product category and rating level	Chi-Square Test
To compare average ratings between high- and low-discount groups	T-Test (Welch)
To compare ratings across different price groups	One-way ANOVA
To assess the normality of the rating count distribution	Shapiro-Wilk Test

5. Analysis and Results

5.1 Does the discounted price significantly impact product ratings?

To explore whether the amount a customer pays (after discount) relates to their satisfaction, we used the Spearman Rank Correlation test, which is ideal for identifying monotonic relationships between numerical variables.

The result showed a correlation coefficient of 0.080 and a p-value of 0.0021, which is statistically significant. Although the correlation is weak, this suggests that as discounted price increases, product ratings also tend to increase slightly.

📝 Interpretation:

Customers might associate higher prices (even discounted ones) with better quality, leading to slightly higher ratings — but the relationship is not strong.

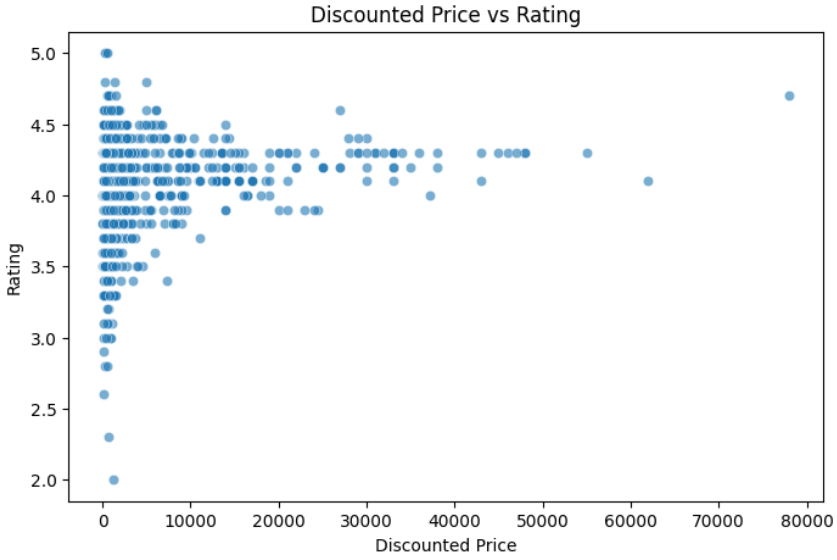

Figure 1: Scatter plot – Discounted Price vs Rating

5.2 Are product categories and rating groups independent?

To understand if a product's category influences whether it receives a high (≥4) or low (<4) rating, we applied the Chi-Square Test of Independence. We grouped ratings into two categories: high and low, and tested their distribution across product categories.

The test returned a Chi-Square statistic of 46.04 with 8 degrees of freedom, and a p-value of 2.33e-07, indicating a strong dependency between product category and rating level.

📝 Interpretation:

This means customers rate some categories more favorably than others — for example, "Office Products" had only high ratings, while "Home & Kitchen" had a mix of both.

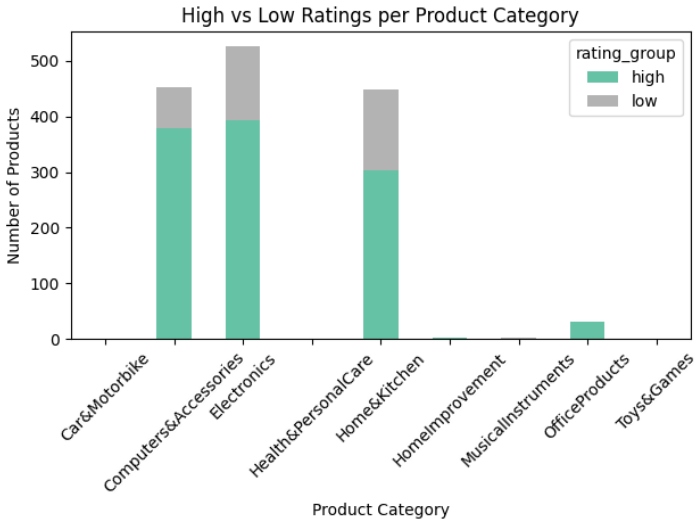

Figure 2: Stacked Bar Chart – High vs Low Ratings per Category

5.3 Is there a significant difference in ratings between high- and low-discount products?

Next, we examined whether deep discounts (≥50%) lead to different ratings compared to smaller discounts (<50%). Using a Welch's T-test (which doesn't assume equal variances), we compared the average ratings of the two groups.

The test yielded a T-statistic of -4.26 and a p-value of 2.17e-05, indicating a statistically significant difference. Interestingly, the negative T-value shows that products with higher discounts actually receive lower ratings on average.

📝 Interpretation:

Heavily discounted items might be perceived as lower quality or as clearance items, which could explain the lower ratings.

5.4 Do more expensive products receive higher ratings?

To analyze if product price impacts rating, we divided products into three price groups: Low (<₹300), Medium (₹300–₹700), and High (>₹700), and performed a One-Way ANOVA to compare their mean ratings.

The test returned an F-statistic of 4.18 and a p-value of 0.0154, suggesting that at least one group differs significantlyfrom the others in terms of average rating.

📝 Interpretation:

While there's a significant difference, the boxplot shows that high prices do not necessarily guarantee better ratings. Medium- and low-priced products perform competitively.

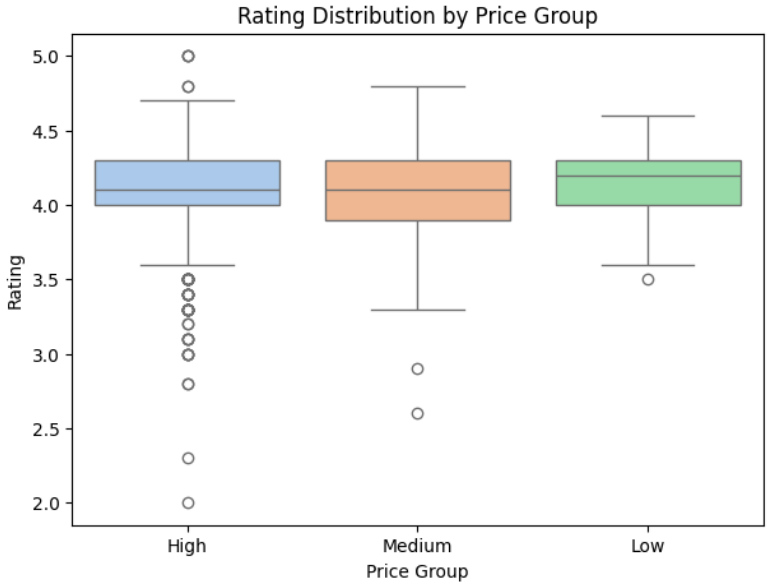

Figure 3: Boxplot – Rating Distribution by Price Group

5.5 Does the distribution of rating counts follow a normal distribution?

Before choosing some statistical tests, it's important to know whether certain variables follow a normal distribution. We used the Shapiro-Wilk Test to assess this for rating_count.

The result was a statistic of 0.41 and a p-value close to 0, indicating a strong deviation from normality.

📝 Interpretation:

Most products have a small number of ratings, while a few have extremely high counts. This "long-tail" distribution is common in online marketplaces, where a handful of products dominate attention.

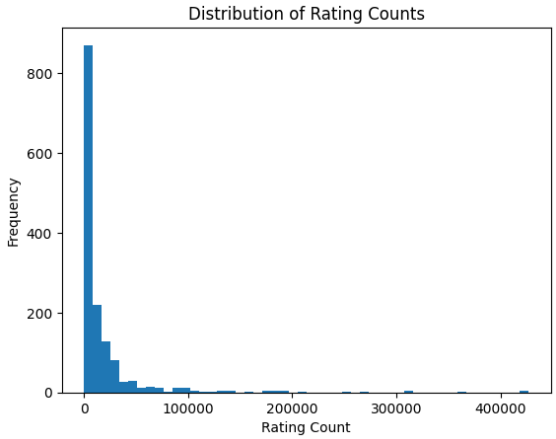

Figure 4: Histogram – Distribution of Rating Counts

6. Conclusion

The analysis revealed that discounts and prices have a measurable effect on ratings, though not always in expected directions. Product category plays a key role in user perception. Most products receive few ratings, with a small number dominating in popularity. These findings support the use of non-parametric tests and suggest strategies for better pricing and promotion on online platforms.