Amazon Sales Analysis

Saba Madadi

March 2025

Abstract

In Amazon sales Analysis, different factors such as pricing details, discount offers, and product types can have an impact. By examining how various features affect sales, it is possible to improve sales strategies and increase business profits. In this study, after analyzing the dataset and taking necessary steps, important insights were gained regarding how these features influence sales. These findings can be useful for enhancing performance.

I. Introduction

Statistical tests were developed in early 20th as essential tools for analyzing data and extracting significant insights. They help researchers to make generalizations from smaller samples and evaluate their hypotheses. Amazon sales analysis using these tests can help business owners improve their performance and increase profits. Analyzing a dataset containing products information is crucial for boosting sales, gaining profits, and staying ahead of competitors. In this study, tests such as the Chi-square test, which examines the impact of two categorical variables, the Spearman Rank Correlation test for measuring strength and direction of association between two ranked variables and ANOVA for comparing means among three or more groups to determine if at least one group mean is significantly different from the others, have been used. By reviewing the results of these tests, insights will gain about between different columns. In the second part, preprocessing of the dataset and its analysis is discussed to gain a deeper understanding of the data. The third section focuses on the statistical tests and their analysis, with the results presented in the fourth section. Finally, the coclusion part summarizes effective affects on saling products in Amazon.

II. Data Exploration

a. Dataset

The dataset has 1,465 rows and 16 columns. Columns include ID, Name, Category, Discounted Price, Actual Price, Discount Percentage, Rating, Rating

Count, About products and also ID, Name for users. There are also Review ID, Review Title, Review Content and Image and Product LIndian rupeeinks. All columns have Object datatype which will be handled on next part.

b. EDA

Initially, by checking for duplicate data, it was found that dataset does not have any duplicates. By looking at a few rows from the beginning and the end, and, a general understanding of dataset was obtained. Next, checking for missing values revealed that there were 2 missing values in the Rating Count column, which were dropped, because there are only 2 missing values and dropping will be fine. To determine data types for each column, first, columns used in statistical tests were identified, which include: discounted_price, rating, category, discount_percentage, actual_price, and rating_count. Next, data types for these columns were selected. The category column was converted to categorical data type. For discounted_price and actual_price, symbols ('Indian Rupee Sign' and ',') were removed, and their data types were converted to int. The discount_percentage column also converted to float by removing the '%' symbol, and rating_count column was converted to float by removing ','. The rating column was converted to float by removing non-numeric values. There was an issue with row 1279 in the dataset, which returned a NaN value. After checking the dataset, it was found that value for this row in rating column is 4.1. This was addressed before completing EDA to prevent additional missing values. At the end of this section, relevant columns for next part were copied into a new DataFrame (df). The columns and their data types are shown in the table below:

Column	**Data Type**
discounted_price	int64
rating	float64
category	category
discount_percentage	float64
actual_price	int64
rating_count	int64

Table 1: df Columns and their Data Types

III. Statistical Analysis

First, the Spearman Rank Correlation Test is used to check if the discounted price significantly impacts product ratings. This test shows that there is a correlation between discounted price and ratings. The plot related to this analysis is shown in the figure below:

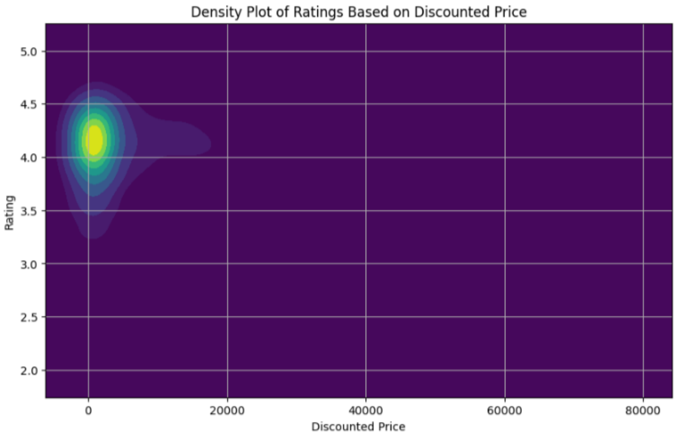

Figure 1: Discounted Price vs. Rating

Next, to better examine the outliers in ‘discounted_price‘, the Interquartile Range (IQR) was calculated, and 51 outliers were removed from dataset. After performing the Spearman Rank Correlation Test again, it was found that there is no correlation between discounted price and ratings. However, the p-value is much closer to alpha, indicating that this result is not very reliable. Therefore, it can be said that the result from first test is more credible. The related plot is shown in figure below:

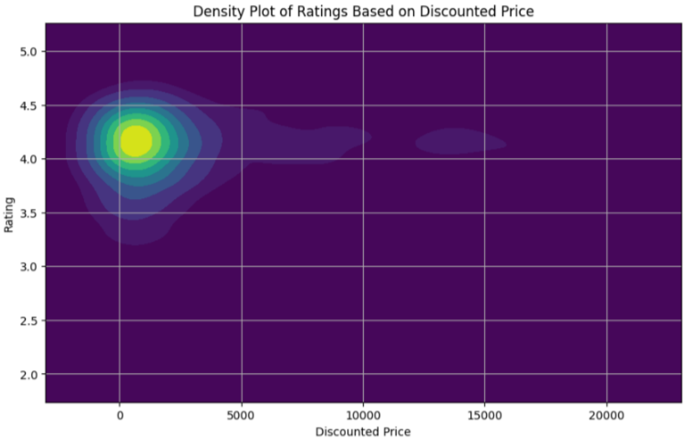

Figure 2: Discounted Price vs. Rating after Removing Outliers

Spearman Rank Correlation Test was performed using the formula:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where d_i is the difference between the ranks of corresponding values, and n is the number of observations.

To investigate whether product categories and high/low ratings are independent, a Chi-Square Test for Independence was conducted. The goal was to determine if there is a significant relationship between product category and rating, with high ratings defined as ratings of 4 or higher, and low ratings defined as ratings below 4.

The Chi-Square statistic is calculated using the formula:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where: - O_{ij} is the observed frequency, - E_{ij} is the expected frequency.

The results of the Chi-Square test indicated that there is a significant relationship between product category and rating.

To investigate whether there is a significant difference in ratings between high-discount and low-discount products, the dataset was split into two groups: products with a discount percentage above 50% and those below this threshold.

An Independent Samples T-test was performed to compare the mean ratings using the formula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where: - $\bar{X}_i$ is the mean rating of each group, - s_i^2 is the variance, - n_i is the sample size.

The mean rating for high discount products was calculated as: 4.063059163059163

The mean rating for low discount products was calculated as: 4.127012987012987

The number of high discount products was: 693

The number of low discount products was: 770

The results of the T-test indicated that there is a significant difference in ratings between high-discount and low-discount products. The plot shown below illustrates the distribution of ratings for the two groups, High Discount and Low Discount.

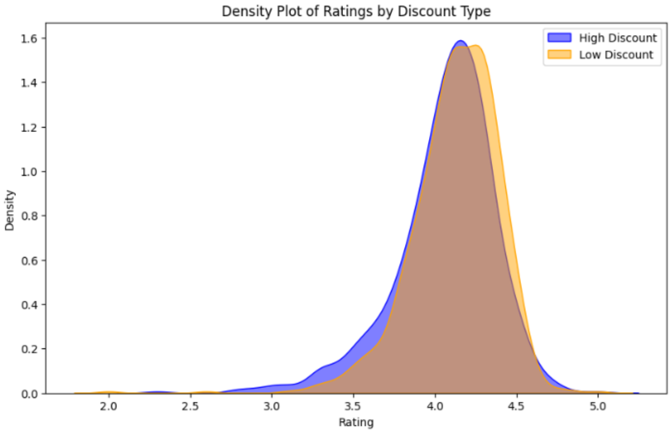

Figure 3: Density Plot of Ratings by Discount Type

To determine if more expensive products receive higher ratings on average, an ANOVA (Analysis of Variance) test was used to check if actual price significantly affects ratings. Products were divided into price groups: 'Low', 'Medium', and 'High'.

The formula for the F statistic is given by:

$$F = \frac{\text{Between-group variance}}{\text{Within-group variance}}$$

Because ANOVA is sensitive to outliers, the following steps were taken before performing the test: 1. Define the outlier limits. 2. Identify outliers. 3. Remove outliers from the dataset.

After categorizing the actual prices into the three specified groups, the ANOVA test was conducted. The results indicated that there is no significant difference in ratings across the price groups. The bar plot shown in the figure below illustrates this conclusion.

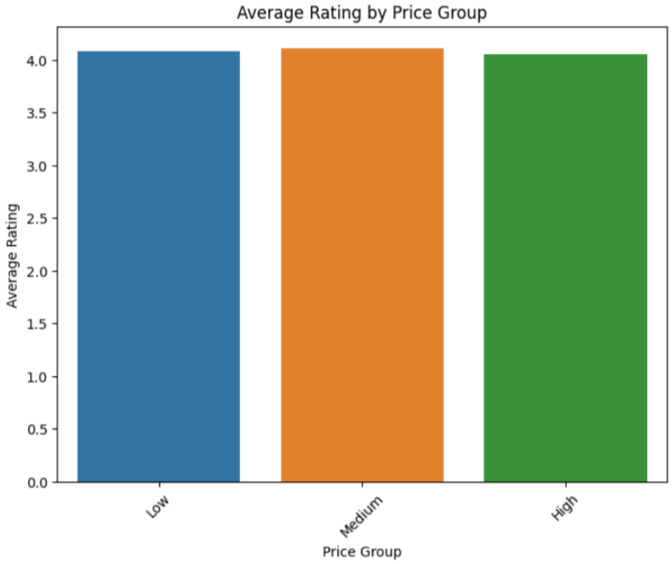

Figure 4: Average Rating by Price Group

To determine if the distribution of rating counts follows a normal distribution, the Kolmogorov-Smirnov test or the Shapiro-Wilk test can be conducted. These tests assess whether the rating count variable adheres to a normal distribution. Before using these tests on dataset removing outliers was handled.

For the Kolmogorov-Smirnov test, the test statistic D_n is defined as follows:

$$D_n = \sup_x |F_n(x) - F(x)|$$

where $F_n(x)$ is the empirical cumulative distribution function (ECDF) of the sample, and $F(x)$ is the cumulative distribution function (CDF) of the normal distribution.

Shapiro-Wilk Test:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where:

- W is the test statistic.
- n is the number of observations.
- a_i are the coefficients calculated from the expected values of the order statistics of a normal distribution.
- $x_{(i)}$ represents the ordered sample values.
- $\bar{x}$ is the mean of the sample.

Both tests indicate that the distribution of rating counts does not follow a normal distribution. The figure blows indicates rating counts distribution vs. normal distribution:

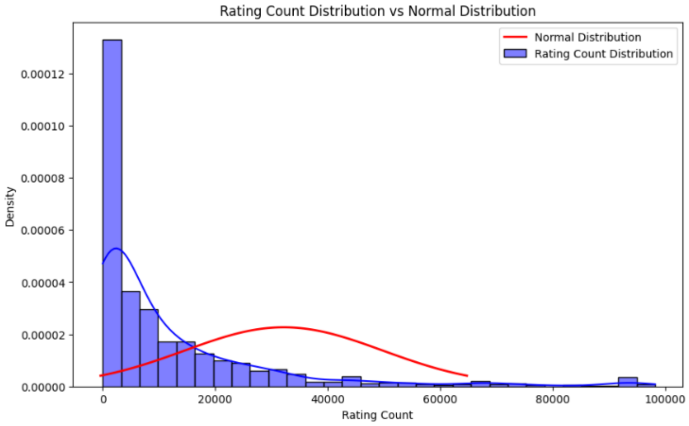

Figure 5: Rating Count Distribution

IV. Results

In this section, the results of the tests conducted earlier will be reviewed and analyzed.

The results of first statistical test (Spearman Rank Correlation Test) showed that, due to p-value being smaller than alpha level of 0.05, there was a correlation between discounted price and ratings. In second phase, after removing outliers at a rate of 0.1 and repeating test, it became evident that, because p-value was larger than alpha, there was no correlation between discounted price and ratings. However, considering that p-value of 0.06 in second phase was close to alpha of 0.05, it can be concluded that results from the first phase are credible.

Table 2: Spearman Rank Correlation Test Results Before Omitting Outliers

Spearman Correlation Coefficient	**P-value**
7.99×10^{-2}	2.21×10^{-3}

Table 3: Spearman Rank Correlation Test Results After Omitting Outliers

Spearman Correlation Coefficient	**P-value**
4.99×10^{-2}	6.04×10^{-2}

To investigate whether product categories and high/low ratings are independent, a Chi-Square test was used. The results indicated that, because p-value was smaller than alpha, there was a significant relationship between product category and rating. The results of this test can be seen in table below:

Table 4: Chi-Square Test Results

Statistic	**Value**
Chi-Square Statistic	385.10
P-value	1.94×10^{-12}
Degrees of Freedom	210

To examine whether there is a significant difference in ratings between high-discount and low-discount products, an Independent Samples T-test was conducted. The results indicated that, because p-value was smaller than alpha, there was a difference in ratings between high-discount and low-discount products.

Table 5: Independent Samples T-test

T-statistic	**P-value**
-4.21×10^{0}	2.66×10^{-5}

To investigate whether more expensive products receive higher ratings on average, an ANOVA test was conducted. The results showed that, because the p-value was smaller than alpha, there was no significant difference in ratings across price groups.

Table 6: ANOVA (Analysis of Variance) Test

F-statistic	**P-value**
4.33×10^{-1}	6.49×10^{-1}

Finally, to determine whether the distribution of rating counts follows a normal distribution, both the Kolmogorov-Smirnov Test and the Shapiro-Wilk Test were conducted. The results indicated that, due to the p-value being smaller than alpha, the rating counts did not follow a normal distribution.

Table 7: Kolmogorov-Smirnov Test

Statistic	**P-value**
2.49×10^{-1}	1.02×10^{-77}

Table 8: Shapiro-Wilk Test

Statistic	**P-value**
4.14×10^{-1}	4.34×10^{-56}

V. Conclusion

Interesting results were obtained from Amazon Sales Analysis using statistical tests. Initially, analysis of whether discounted price significantly impacted product ratings showed that a relationship existed between ratings and discounted prices. It was concluded that amount of discount given to customers affected ratings they provided. Next, in examination of whether product categories and high/low ratings were independent, it was found that high or low ratings were related to type of category. Additionally, there was a significant difference in ratings between high-discount and low-discount products. Furthermore, analysis of whether more expensive products received higher ratings on average revealed that higher-priced products did not receive higher ratings, and in all price ranges, the given rating was 4. Finally, investigation into distribution of rating counts indicated that this distribution was not normal. For future work, analysis could continue with other columns and utilize more statistical tests. For example, it could be explored whether giving discounts to all products at different prices would still attract customers.