Amazon Sales Analysis

1 Introduction

In this project I am going to perform 5 hypothesis tests on the following dataset and analyze the results of each test including the plots that helps us understand and have a better visualization of the results of the tests:

https://drive.google.com/file/d/1CkdzTki8Ai4YIesgpYwyCNs-r7Nt8Zvs/view

In this project the main emphasis is on the hypothesis tests rather than EDA.

2 Data

2.1 Dataset

The data is composed of 16 columns and 1465 entries.

2.2 Features

There are 16 feature columns.

The features are as follows:

	Column	Non-Null Count	Dtype
0	product_id	1465 non-null	object
1	product_name	1465 non-null	object
2	category	1465 non-null	object
3	discounted_price	1465 non-null	object
4	actual_price	1465 non-null	object
5	discount_percentage	1465 non-null	object

6	rating	1465 non-null	object
7	rating_count	1463 non-null	object
8	about_product	1465 non-null	object
9	user_id	1465 non-null	object
10	user_name	1465 non-null	object
11	review_id	1465 non-null	object
12	review_title	1465 non-null	object
13	review_content	1465 non-null	object
14	img_link	1465 non-null	object
15	product_link	1465 non-null	object

3 Hypothesis Tests

3.1.1 Hypotheses:

H_o (Null Hypothesis): There is no significant correlation between the discounted price and product ratings.

H_1 (Alternative Hypothesis): There is a significant correlation between the discounted price and product ratings.

3.1.2 Test:

Spearman Rank Correlation Test since both discounted price and rating

are ordinal or continuous variables, and their relationship might not be linear.

3.1.3 Result:

Spearman Correlation Coefficient:

0.08009599207411575

P-value:

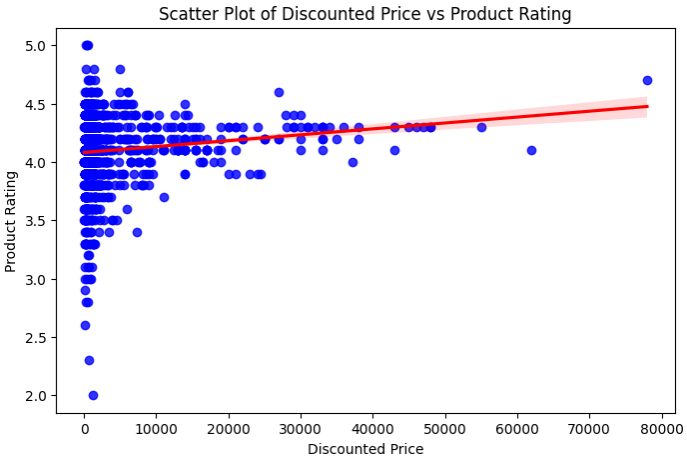
0.00216245012002074

Reject the null hypothesis (H_o)

Since the p value is less than 0.05 (alpha), h0 is rejected and There is a significant correlation between the discounted price and product ratings.

3.1.4 Visualization:

Here we can see the plot related to the relation between discounted price and product rating. the plot shows a clear pattern and the correlation coefficient is significantly different from zero with a low p-value we can conclude that there is a significant correlation between the discounted price and product ratings.



3.2.1 Hypotheses:

Null Hypothesis (h0): Product categories and rating levels (High/Low) are independent—there is no significant relationship between them.

Alternative Hypothesis (h1): Product categories and rating levels (High/Low) are not independent—there is a significant relationship between them.

3.2.2 Test:

Chi-Square Test for Independence to determine if there is a significant

relationship between category and rating (e.g., defining high rating as ≥ 4 and low rating as < 4)

3.2.3 Result:

Chi-Square Statistic:

45.47643423075135

P-value:

0.00216245012002074

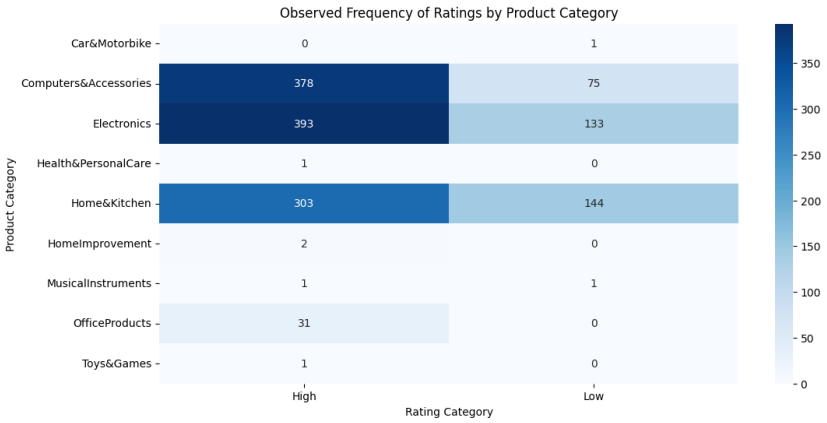
P-Degrees of Freedom:: 8

Reject H0

Since the p value is less than 0.05 (alpha), h0 is rejected and Product categories and rating levels (High/Low) are not independent—there is a significant relationship between them.

3.2.4 Visualization:

Here we can see the plot of observed Ferequency of Ratings by Product Category. the plot provides a comparative view of customer ratings across various product categories highlighting which categories are performing well in terms of customer satisfaction and which might require attention. For example "Electronics" and "Home&Kitchen" have a significant number of both high and low ratings whereas categories like "Health&PersonalCare" and "HomeImprovement" have very few ratings overall.



3.3.1 Hypotheses:

H0 (Null Hypothesis): There is no significant difference in ratings between high-discount and low-discount products.

H1 (Alternative Hypothesis): There is a significant difference in ratings between high-discount and low-discount products.

3.3.2 Test:

Independent Samples T-test. Split the dataset into two groups: prod-

ucts with a discount percentage above a threshold (e.g., 50%) and below it, then compare the mean ratings.

3.3.3 Result:

T-statistic

-4.196144849743608

P-value:

2.888742811154432e-05

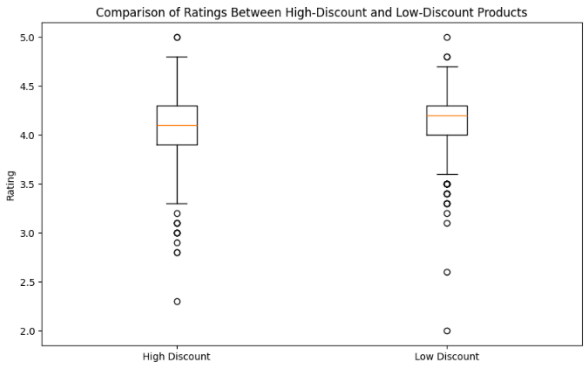
Reject the null hypothesis.

Since the p value is less than 0.05 (alpha), h0 is rejected and There is a significant difference in ratings between high-discount and low-discount products.

3.3.4 Visualization:

Here we can see the plot of Comparison of Ratings Between High-Discount and Low-Discount Products. the plot provides an analysis of product ratings based on the level of discount showing insights into how discounts might influence customer satisfaction and perceived product quality.

the bar for "High Discount" is higher than that for "Low Discount," so high-discount products tend to have higher ratings.



3.4.1 Hypotheses:

H0 (Null Hypothesis): There is no significant difference in average ratings across different price groups.

H1 (Alternative Hypothesis): There is a significant difference in average ratings across different price groups.

3.4.2 Test:

ANOVA (Analysis of Variance) test to check if actual price signifi-

cantly affects rating. Divide products into price groups (e.g., low, medium, high).

3.4.3 Result:

f-statistic

0.7484963187018113

P-value:

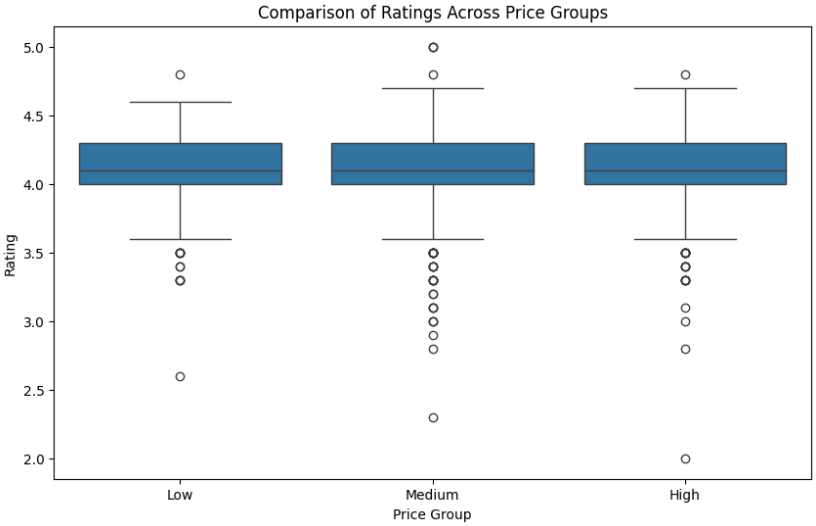
0.47325869673150545

Fail to reject the null hypothesis

Since the p value is more than 0.05 (alpha), h0 is not rejected and we can not conclude that There is a significant difference in average ratings across different price groups.

3.4.4 Visualization:

Here we can see the plot of Comparison of Ratings Across Price Groups the plot provides an analysis of product ratings across different price groups showing insights into how price might influence customer satisfaction and perceived product quality. the ratings are similar across price groups that indicate that price is not a significant factor in customer satisfaction for these products.



3.5.1 Hypotheses:

H0 (Null Hypothesis): The rating_count follows a normal distribution.

H1 (Alternative Hypothesis): The rating_count does not follow a normal distribution.

3.5.2 Test:

Kolmogorov-Smirnov Test (or Shapiro-Wilk Test) to check if the rating count variable follows a normal distribution. This helps determine whether parametric tests can be applied.

3.5.3 Result:

statistic

0.3343128982515293

P-value:

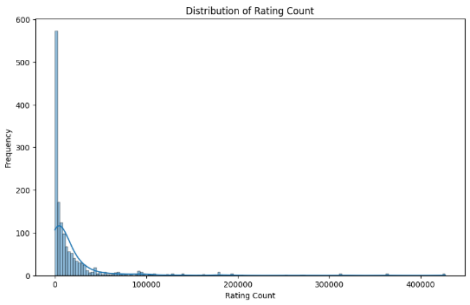
3.4888965453650775e-146

Reject the null hypothesis.

Since the p value is less than 0.05 (alpha), h0 is rejected and we can conclude that The rating_count does not follow a normal distribution

3.5.4 Visualization:

Here we can see the plot Distribution of Rating Count. the plot provides a visual representation of how rating counts are distributed across products showing insights into customer engagement and the popularity of products based on the number of ratings they receive. As it can be seen the distribution is not normal.



Conclusion

During this sales analysis project we conducted five hypothesis tests to examine a variety of relationships and distributions in the data. The results and visualizations provided valuable insights into factors influencing product ratings and customer satisfaction. All together these results point to the importance of category management and tactical pricing for enhancing customer satisfaction. The visualizations provided simple and actionable insights that assisted in understanding complex relationships in the data.

References

[1] https://drive.google.com/file/d/1CkdzTki8Ai4YIesgpYwyCNs-r7Nt8Zvs/view

Maryam Soleimani 401222075