UBFC

UNIVERSITÉ
BOURGOGNE FRANCHE-COMTÉ

Topic: An Analysis of the Impact of Ratings and Reviews on Amazon Sales Across Categories

Course: Data Analysis and Software (R Programming)

Report By: Liza Gyamaa Oppong

Contents

1. Introduction 3

2. Methodology 4

2.1. Data Source 4

2.2. Data Preparation 4

3. Analysis Methods 4

4. Results 5

4.1. Correlation Analysis 5

4.2. Simple Linear Regression: 6

4.3. Multiple Linear Regression: 6

4.4. Visualizations 8

4.4.1. Scatterplot: 8

4.4.2. Bar Chart: 8

4.2.3. Boxplot: 9

5. Summary of Findings and Discussion 10

6. Conclusion 10

7. References 10

Abstract

This study explores how ratings and reviews impact sales on Amazon across various product categories. Using a dataset sourced from Kaggle, a sales metric was estimated by combining actual prices and ratings. Correlation and regression analysis were conducted to identify trends and predictors. The findings reveal that while higher ratings correlate with increased sales, pricing plays a more dominant role. Review quantity alone has minimal impact, highlighting the importance of quality feedback. The results offer actionable insights for sellers aiming to optimize their strategies in e-commerce.

1. Introduction

Ratings and reviews are crucial in how people shop online, shaping decisions and influencing what ends up in their carts. For Amazon sellers, figuring out how these factors impact sales can unlock better strategies for products and promotions. As someone who frequently shops online, I find myself consistently relying on ratings and reviews to make informed purchase decisions. That's what sparked my interest to explore just how much ratings and reviews affect sales on Amazon, one of the biggest online marketplaces in the world. This report investigates: **"How ratings and reviews impact Amazon sales across categories."**

The topic is significant for sellers aiming to optimize pricing and promotional strategies and for consumers seeking quality products. By analyzing this relationship, we gain insights into consumer behavior, providing valuable data for improving online retail strategy. The dataset used for this analysis was sourced from Kaggle and contains detailed information about product prices, discounts, ratings, reviews, categories, ratings count etc. Since the dataset did not include a sales column, I estimated it by multiplying the actual price by its rating and dividing it by 5. This assumes that highly rated, more expensive items generally sell better.

This generated column provides a proxy for estimating sales performance based on price and customer ratings. The inclusion of this column is crucial, as it enables us to quantitatively analyze the relationship between ratings, reviews, and sales.

This research aims to:

1. Identify how product ratings influence sales performance on Amazon across different categories?
2. Examine the relationship between the number of customer reviews and sales on Amazon?
3. How significant is the role of pricing in driving sales compared to ratings and reviews?
4. Do certain product categories experience a stronger correlation between ratings, reviews, and sales than others?

2. Methodology

2.1. Data Source

The dataset was obtained from Kaggle and consisted of product information. Since the dataset lacked a sales column, this was generated.

This method assumes that higher ratings directly translate into greater sales, moderated by product price. The choice of formula reflects the assumption that ratings serve as a proxy for consumer trust, and higher-priced items typically have greater potential sales.

To ensure the dataset was ready for analysis, several additional steps were taken to address the complexities of deriving meaningful insights from real-world data. By focusing on the integrity and relevance of each variable, the methodology prioritizes a structured and transparent approach to data transformation.

2.2. Data Preparation

The dataset underwent several preprocessing steps:

- Handling Missing Values: Rows with missing data were removed.
- Data Type Conversion: Key columns such as sales, rating, and discount percentage were converted to numeric types.
- Column Selection: Only relevant columns were retained for analysis: discounted price, actual price, discount percentage, rating, rating count, sales, review title, and category.
- Exploratory Data Analysis: Patterns in the data were visualized to understand relationships and identify anomalies.
- Sales Column Derivation: The sales column was generated to serve as a proxy for actual sales data, which was unavailable. This involved calculating the product's actual price and rating, divided by 5. This column allowed the analysis to quantitatively link ratings and pricing to estimated sales, enabling meaningful comparisons and predictions.
- Validation: The generated sales data were checked against expected trends to ensure that outliers or anomalies did not skew the results.

3. Analysis Methods

This phase of the methodology combined statistical and graphical tools to uncover trends and evaluate hypotheses. Key steps included:

- **Descriptive Statistics:** Summarized sales, ratings, and review counts.
- **Correlation Analysis:** Explored relationships between sales, ratings, and review counts.
- **Regression Models:** Implemented linear regression models to identify predictors of sales.
- **Category Comparison:** Performed hypothesis testing (t-tests) to evaluate differences in sales across key categories.

- **Visualizations:** Scatterplots, bar charts, bar plot and boxplots were created to visualize key trends.

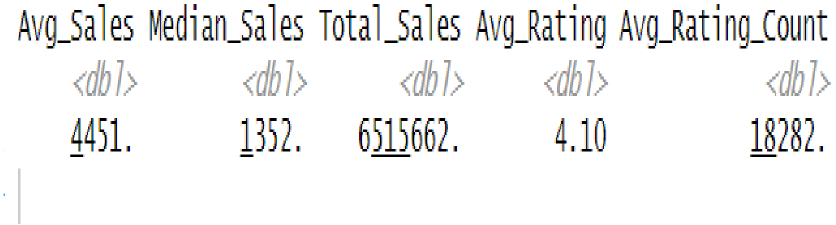
4. Results

Correlation Analysis and Descriptive Statistics

Key statistics revealed insights into sales, ratings, and review counts:

- **Average Sales:** $4451
- **Median Sales:** $1352
- **Total Sales:** $6,515,662
- **Average Rating:** 4.10
- **Average Rating Count:** 18,282

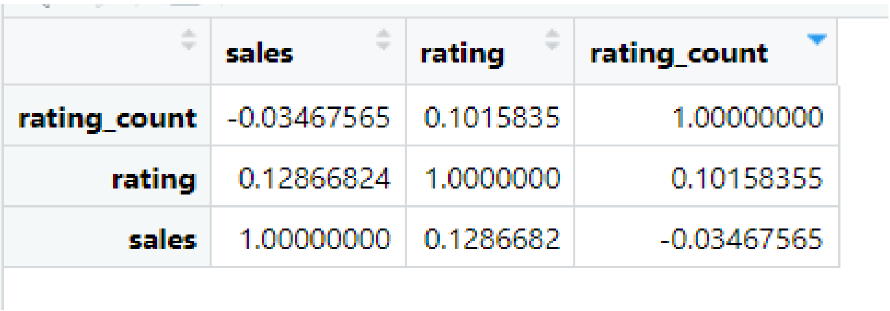
These findings suggest a concentration of sales around a few high-performing products, reflecting typical e-commerce dynamics.



```
Avg_Sales Median_Sales Total_Sales Avg_Rating Avg_Rating_Count
    <dbl>        <dbl>       <dbl>      <dbl>            <dbl>
    4451.        1352.    6515662.       4.10           18282.
```

4.1. Correlation Analysis

Correlation measures the degree to which two variables move together. In this analysis, a weak positive correlation (ρ = 0.1287) was found between ratings and sales. This indicates that while higher ratings are associated with increased sales, the relationship is not strong enough to suggest a direct or dominant influence. Several external factors, such as product visibility, pricing strategies, and advertising, could dilute this relationship. On the other hand, review count displayed an almost negligible correlation with sales (ρ = -0.0347), suggesting that merely accumulating reviews without achieving high ratings does not significantly boost sales.



	sales	rating	rating_count
rating_count	-0.03467565	0.1015835	1.00000000
rating	0.12866824	1.0000000	0.10158355
sales	1.00000000	0.1286682	-0.03467565

The weak correlations suggest that ratings and reviews alone are insufficient to predict sales accurately, and their impact might vary significantly across different product categories and price ranges. Regression Analysis and Regression models were employed to assess the combined influence of ratings, review count, and price on sales.

4.2. Simple Linear Regression:

This model analyzed the direct relationship between ratings and sales, revealing that each unit increase in rating corresponded to an average increase of $3846.12 in sales. Although significant, the model's simplicity means it does not account for other influential factors like price and review count.

4.3. Multiple Linear Regression:

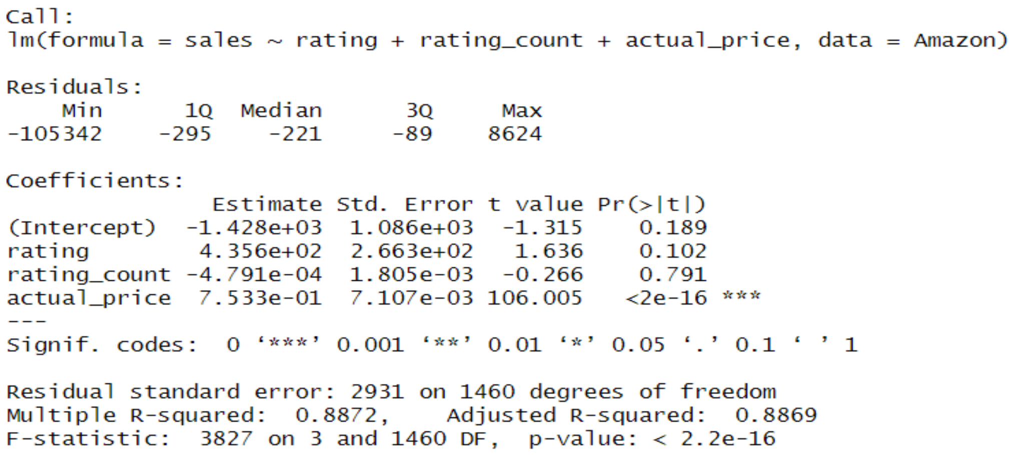
A more comprehensive model incorporating rating, rating count, and actual price demonstrated the following:

Actual Price: A highly significant predictor ($p < 0.001$), reflecting its dominant role in driving sales.

Rating: While ratings positively influenced sales, the effect was weaker ($p = 0.102$), suggesting that their impact is secondary to pricing.

Review Count: Negligible effect ($p = 0.791$), aligning with the earlier correlation analysis.

The Adjusted R-squared value of 0.887 indicates that the model explains approximately 88.7% of the variability in sales, underscoring the strong influence of the included predictors. However, the high residual standard error highlights the variability in sales that may arise from other unaccounted for factors like seasonal trends or product marketing.



```
Call:
lm(formula = sales ~ rating + rating_count + actual_price, data = Amazon)

Residuals:
    Min      1Q  Median      3Q     Max 
-105342    -295    -221     -89    8624 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -1.428e+03  1.086e+03  -1.315    0.189    
rating        4.356e+02  2.663e+02   1.636    0.102    
rating_count -4.791e-04  1.805e-03  -0.266    0.791    
actual_price  7.533e-01  7.107e-03 106.005   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2931 on 1460 degrees of freedom
Multiple R-squared:  0.8872,	Adjusted R-squared:  0.8869 
F-statistic:  3827 on 3 and 1460 DF,  p-value: < 2.2e-16
```

Category Insights

Electronics:

- Average Sales: $8196
- Total Sales: $4,311,219

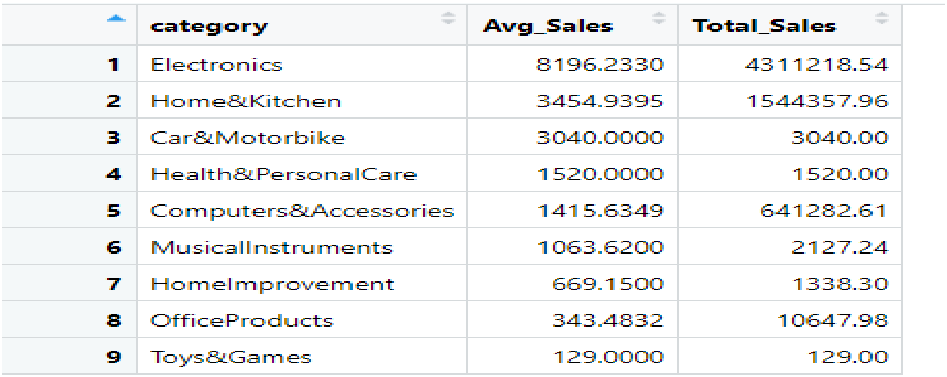
Computers & Accessories:

- Average Sales: $1416

Statistically lower sales than Electronics (Welch’s t-test: τ = -12.237, p < 2.2e-16).

Other Categories:

Varied sales patterns, with Home & Kitchen performing moderately well.

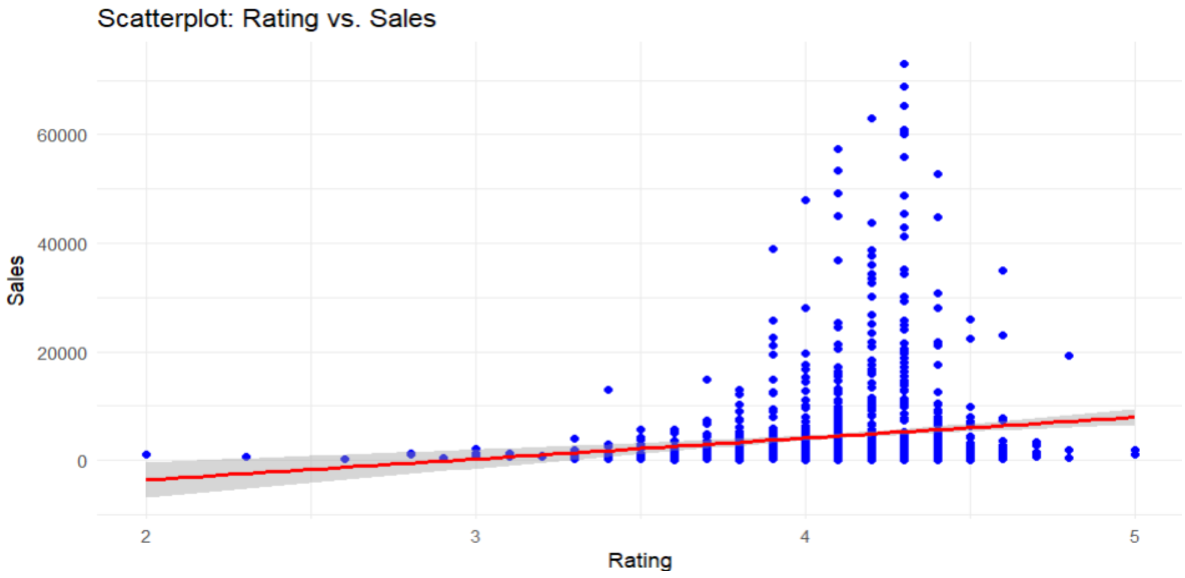


	category	Avg_Sales	Total_Sales
1	Electronics	8196.2330	4311218.54
2	Home&Kitchen	3454.9395	1544357.96
3	Car&Motorbike	3040.0000	3040.00
4	Health&PersonalCare	1520.0000	1520.00
5	Computers&Accessories	1415.6349	641282.61
6	MusicalInstruments	1063.6200	2127.24
7	HomeImprovement	669.1500	1338.30
8	OfficeProducts	343.4832	10647.98
9	Toys&Games	129.0000	129.00

4.4. Visualizations

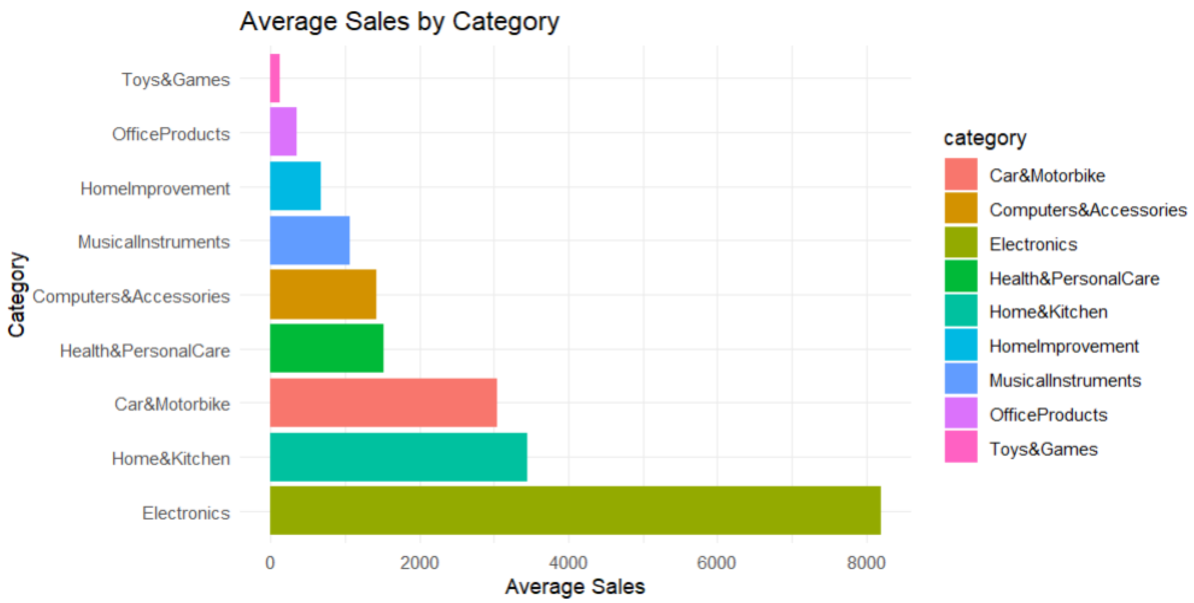
4.4.1. Scatterplot: Ratings vs. Sales: A trendline showed a positive relationship, albeit with significant variability.

This visualization highlights the weak positive correlation between ratings and sales. The spread of data points emphasizes the complexity of factors influencing sales, showcasing that high ratings are not always predictive of high sales.



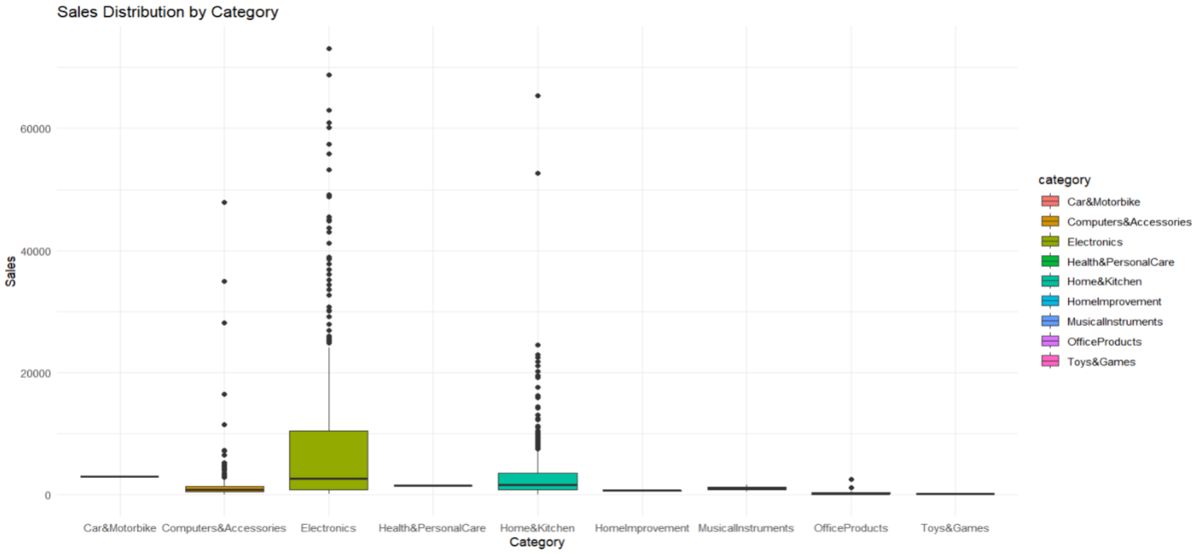
4.4.2. Bar Chart: Average sales by category: Electronics dominated, followed by Home & Kitchen and Computers & Accessories.

It visually demonstrates category-level differences in average sales, aiding in identifying top-performing product segments.



4.2.3. Boxplot: Sales distribution across categories: Electronics showed the highest variability and presence of outliers.

This visualization captures the diversity in sales performance across categories and underscores areas for further investigation.



5. Summary of Findings and Discussion

The analysis reveals that while ratings positively influence sales, the relationship is not strong enough to guarantee success. Pricing plays a more significant role, as shown by its dominant predictive power in the regression model. The negligible impact of review count suggests that quality feedback is more critical than the number of reviews.

Sales variability across categories indicates that consumer behavior is influenced by product type, with Electronics consistently leading in both average and total sales.

Implications:

- Sellers should prioritize competitive pricing while maintaining high ratings to optimize sales.
- Researchers can explore additional factors influencing e-commerce success, such as marketing strategies and seasonal trends.
- Visual tools like scatterplots and bar charts offer clear insights into key dynamics, aiding further investigation and strategy development.

This underscores the importance of balanced strategies and further exploration of external factors to enhance e-commerce performance.

6. Conclusion

The analysis shows that while ratings positively influence sales, pricing is the most critical factor driving them. Electronics outperformed other categories in average and total sales. The weak correlation between review count and sales highlights that quality feedback matters more than quantity. Variability in sales across categories suggests external factors like marketing, brand reputation, and seasonal demand play significant roles. The adjusted R-squared value of 0.887 indicates a strong model but leaves room for exploring other variables. A balanced strategy involving competitive pricing, high ratings, and quality reviews is essential. Future research should explore factors like product descriptions and advertising to better understand e-commerce dynamics.

7. References

Kaggle: Dataset Source - Amazon Sales Dataset