Amazon Sales Analysis

Introduction

In the vast landscape of Amazon's marketplace, product sales volume and customer ratings are crucial indicators of product success and consumer satisfaction. This report delves into two essential questions:

1. How do customer ratings vary across Amazon's product categories, and are the ratings related to sales volume?
2. How are Amazon sales and ratings related to its product discount?

Our motivation stems from a desire to understand the underlying influence of discount level on customers' purchasing behaviors and the hidden dynamics that shape customer experiences in e-commerce.

We posit that distinct factors, possibly including price and discount strategies, significantly influence customer purchases and ratings in varying product categories. This thesis stands at the core of our analysis: Customer satisfaction on Amazon is a nuanced tapestry, woven with diverse influences that vary across categories, offering valuable insights into the mechanics of online consumer behavior.

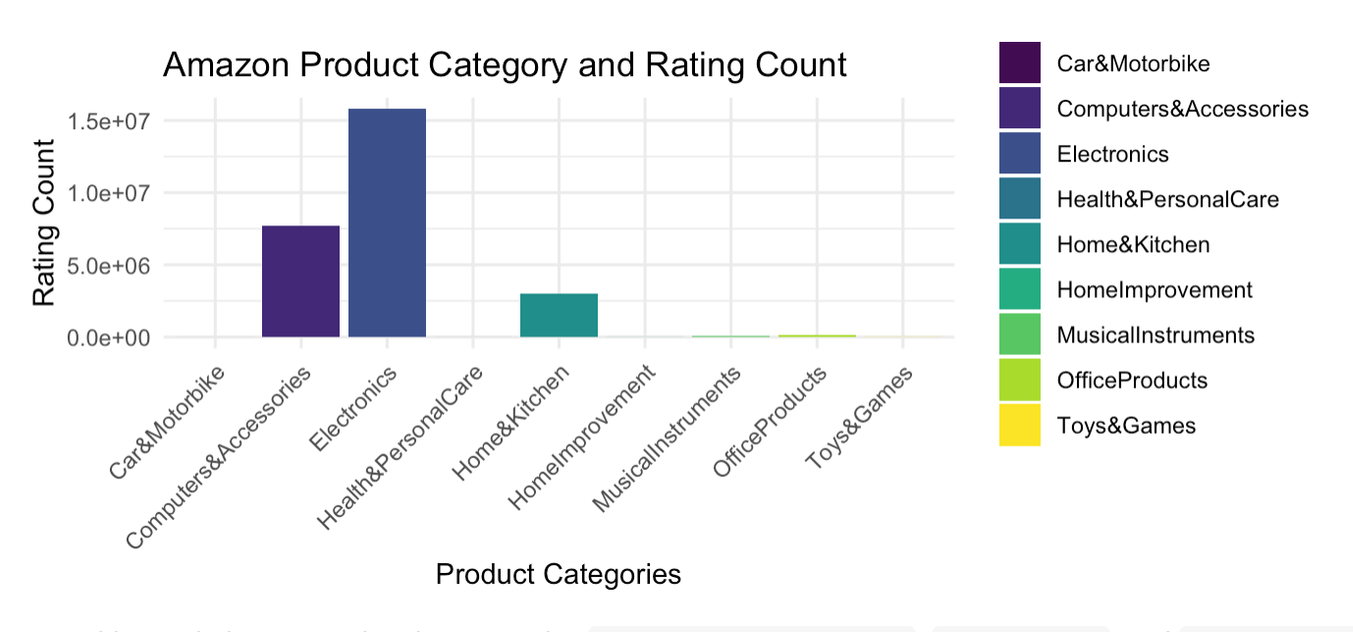
Background

- **About the data:** The dataset under consideration was compiled from Amazon's publicly accessible customer reviews and product information sections. It was scraped in January 2023 from the Official Website of Amazon, and contains data of more than 1,000 Amazon Product's Ratings and Reviews with relative details listed. The data can be considered a sample of the larger population of all Amazon reviews globally, specifically pertaining to the tech accessories category. However, since the dataset contains data gathered from a limited time period, its reliability and the results we get from analysis are limited. As one of the most influential American Tech Multi-National Company, Amazon's sales data is valuable for conducting analysis for both research and business purposes. Some of the products in this dataset has NA values, and we ignore them in this project.
- **Data Citation:** Karkavelraja J. Januart, 2023. "Amazon Sales Dataset". Kaggle. https://www.kaggle.com/datasets/karkavelrajaj/amazon-sales-dataset/data
- **Key Variables:**
  - product_id: Unique identifier for each product.
  - product_name: Name of the product.
  - category: Classification of the product within Amazon's hierarchy.
  - discounted_price: Product's discounted and original prices.
  - actual_price: Product's original price
  - discount_percentage: The percentage reduction from the original price.
  - rating: The average user rating for the product.
  - rating_count: The number of ratings a product has received.
  - review_content: Text content of the user review.

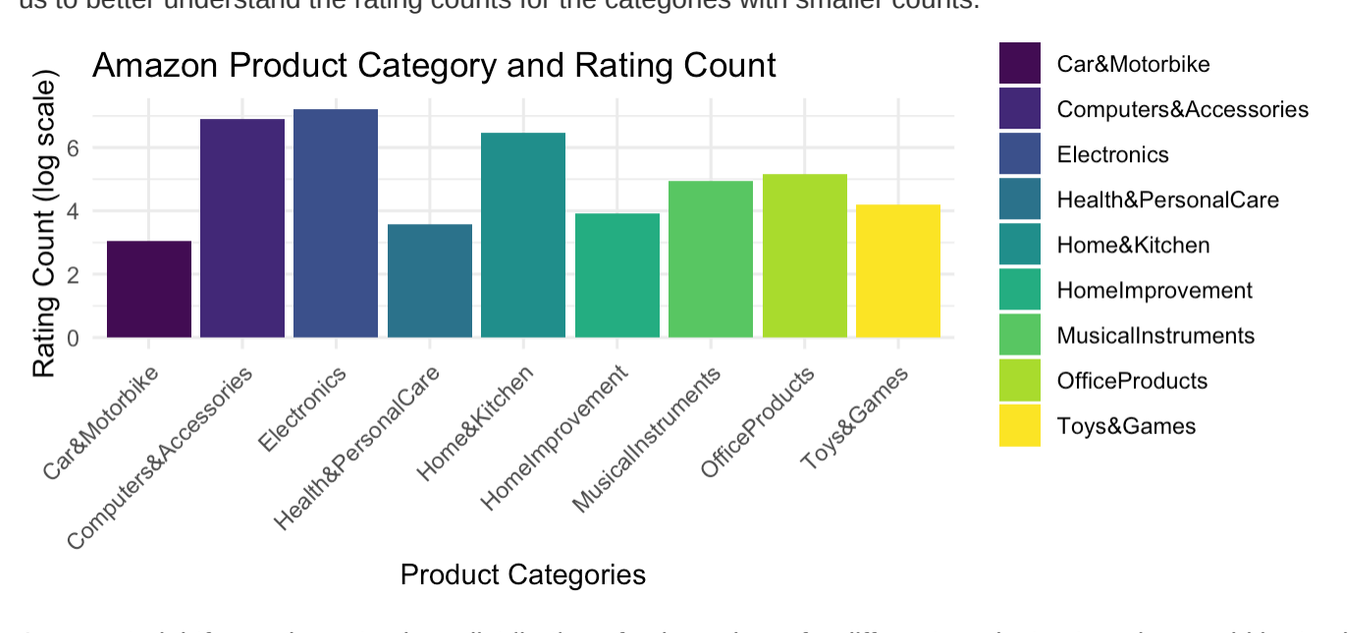
> Our first focus is to explore and analyze the sales and rating performance of diverse product categories, and the correlation between customer ratings and sales volume for all products. Our second focus is to find out the correlation between discount level, sales volume, and ratings.

Analysis - Question 1

> For the first question, we organize the data and focus on the sales distribution for different Amazon product categories. We are assuming that rating counts is a representation of sales volume. This analysis will help us figure out consumer preferences of product types.



From this graph, it appears that the categories `Computers&Accessories`, `Electronics`, and `Home&Kitchen` have most rating counts, or in other words, sales volume. We then use log of `Rating Count` so that the difference in bar heights are not as huge as they are now. This allows us to better understand the rating counts for the categories with smaller counts.

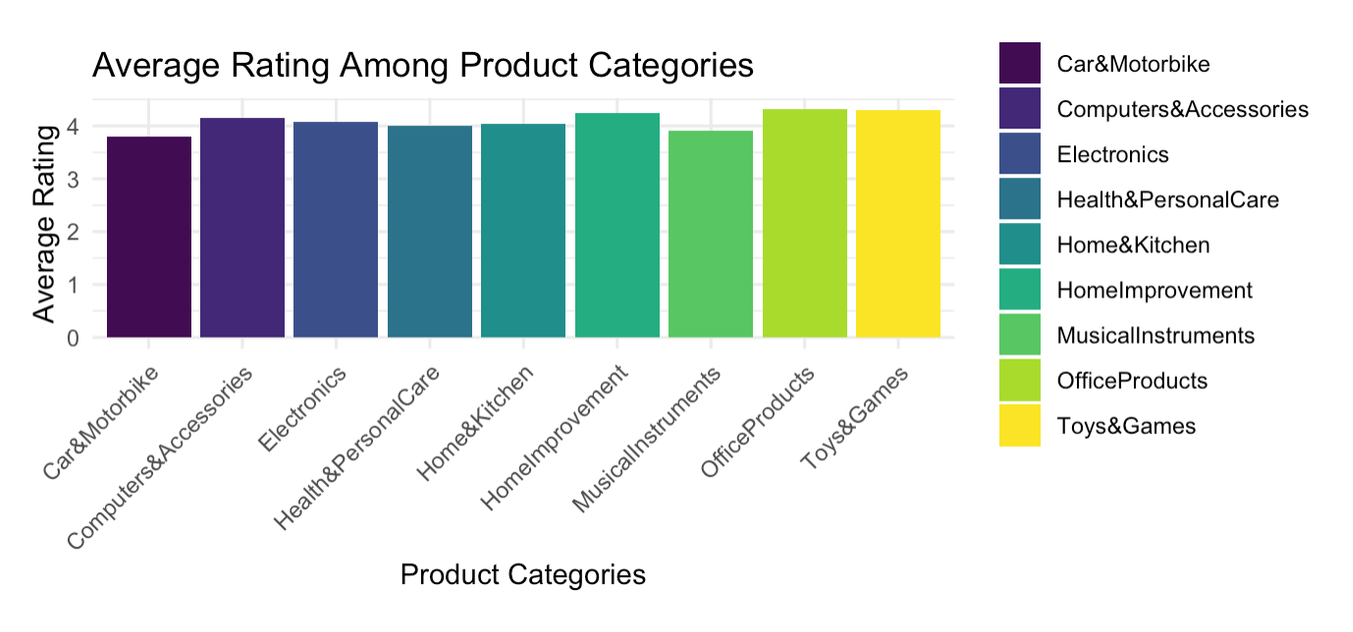


A more straightforward way to show distribution of sales volume for different product categories would be to show the different percentage of total rating counts each category accounts for.

```
## # A tibble: 9 × 2
##   category               percentage
##   <chr>                       <dbl>
## 1 Car&Motorbike           0.0000418
## 2 Computers&Accessories   0.289
## 3 Electronics             0.590
## 4 Health&PersonalCare     0.000137
## 5 Home&Kitchen            0.112
## 6 HomeImprovement         0.000320
## 7 MusicalInstruments      0.00332
## 8 OfficeProducts          0.00559
## 9 Toys&Games              0.000593
```

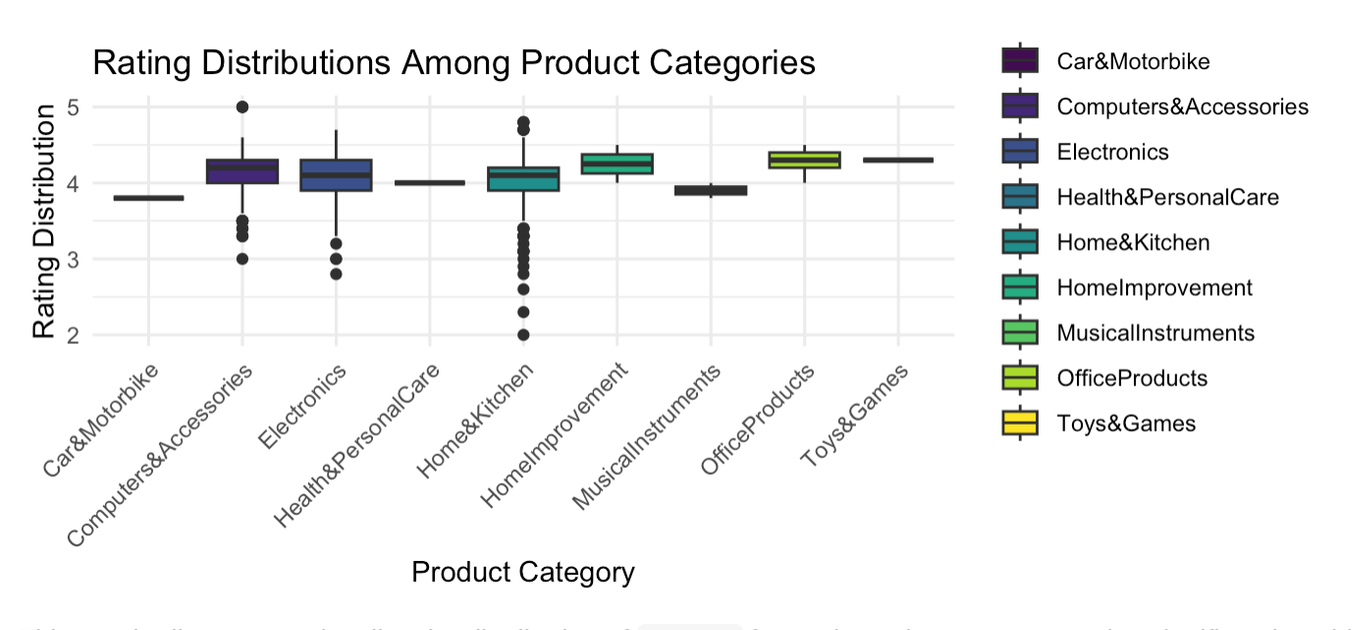
Overall, varying product categories have significantly different sales volume. It is evident that in total, `Computers&Accessories`, `Electronics`, and `Home&Kitchen` occupy a considerably large portion of the total rating count/sales volume for all Amazon products.

> Next, we look at customer ratings for each product category to analyze consumer satiscation level for each category. We will use mean rating as a proxy of rating for each product category to visualize the data.



Here, we can see that the average rating for each category is very similar, which could be an indication for high consumer satifaction level for their purchases on Amazon. It is worth mentioning that since the categories of `Computers&Accessories`, `Electronics`, and `Home&Kitchen` have most ratings data, the mean ratings for these categories would be more accurate indicators for customer satisfaction when compared to the other categories.

> We then dive deeper to see the rating distribution for each product category using box plots.

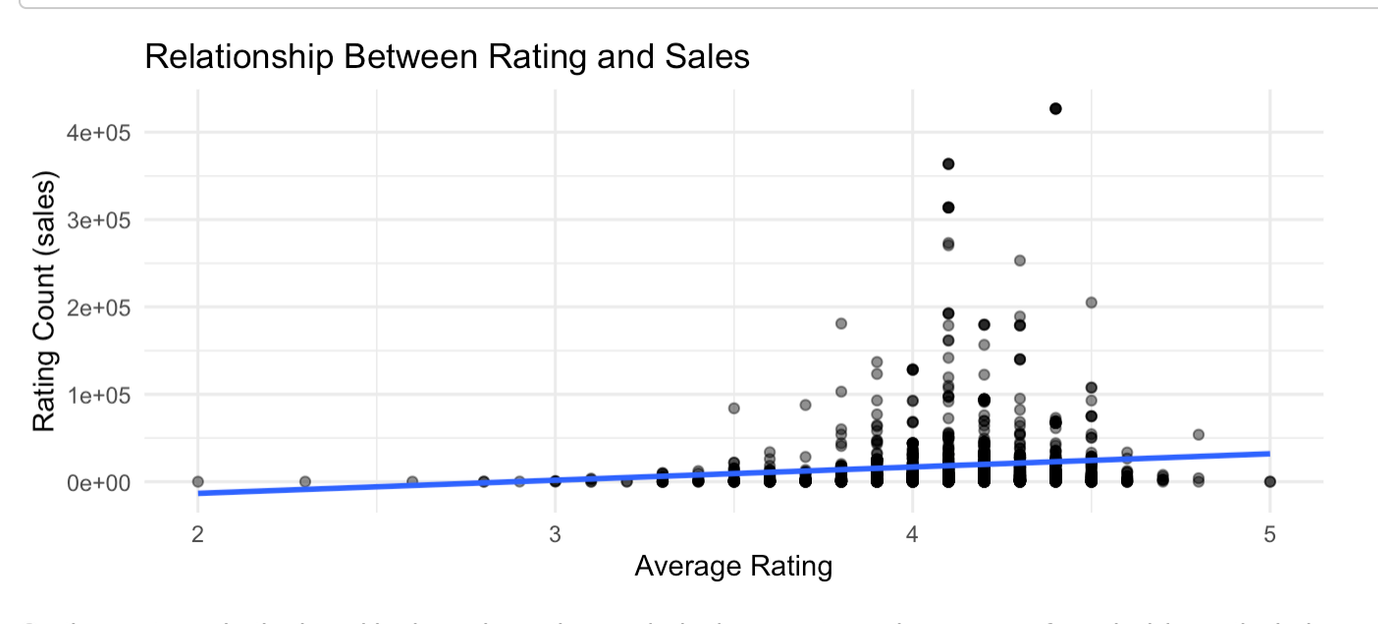


This graph allows us to visualize the distribution of `Rating` for each product category varies significantly. While some categories have wider spreads of ratings distribution, some have really small ones. When comparing this graph to the bar plots mapping `Product Category` against `Rating Count`, we noticed that categories with higher sales volume (rating count) have larger spreads of distribution, which makes sense since the larger the database, the better it presents the features of the data. One thing that is notable is that within all the categories, `Home&Kitchen` has most outliers both over and under its IQR, which indicates that customer satisfaction on products in this category vary most significantly.

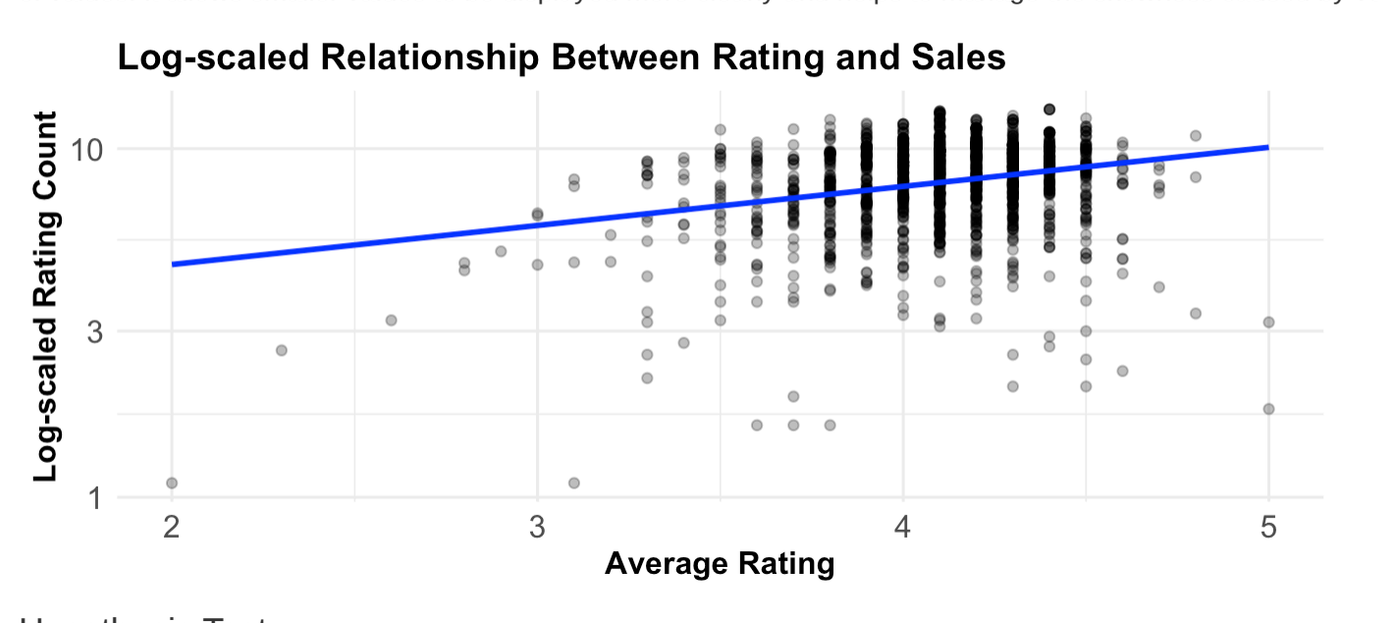
> Next, we're examining whether there's a connection between how customers rate products and the number of sales on Amazon across all product categories. Again, since actual sales figures are not available, we'll use the number of ratings as a stand-in to estimate sales volume. The upcoming graph plots product ratings against the number of ratings to visually assess if higher-rated products also tend to sell more. This simple yet effective analysis will help us understand if there's a trend that suggests customer ratings might influence sales.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   2.000   4.000   4.100   4.097   4.300   5.000       1
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##       2    1186    5179   18296   17336  426973       2
```



On the next graph, the logarithmic scale on the y-axis (rating count, used as a proxy for sales) is particularly useful for datasets with a wide range of values. It allows smaller values to be displayed more clearly and helps to manage the skewness caused by very large values.



Hypothesis Test

$$H_0: \theta = 0$$
$$H_a: \theta \neq 0$$

We test the null that our slope parameter θ is 0, suggesting no correlation between `Average Product Rating` and `Number of Ratings`, against the alternative that there is a correlation.

Test Statistic

$$T = \frac{\hat{\theta} - 0}{s_{\hat{\theta}}}$$

$$s_{\hat{\theta}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

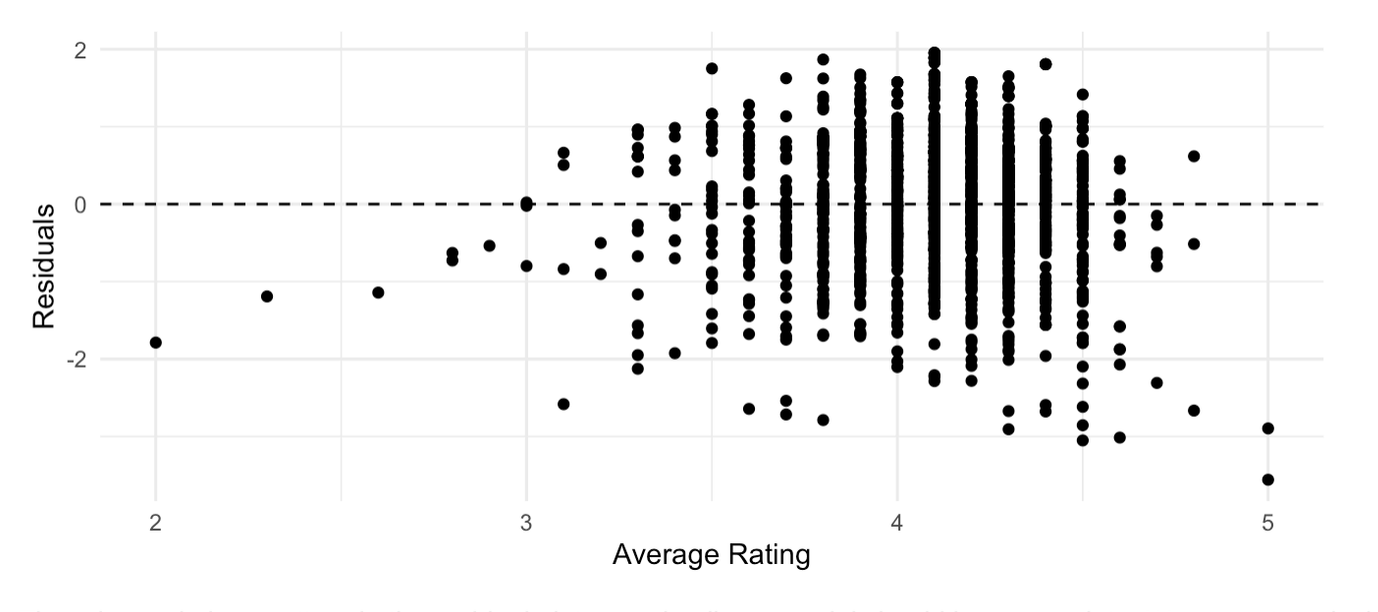
```
## 
## Call:
## lm(formula = log10(rating_count) ~ rating, data = amazon)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.5588 -0.5063  0.0837  0.6243  1.9536 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   0.6432     0.3191   2.016    0.044 *  
## rating        0.7229     0.0777   9.304   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.8597 on 1460 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.05598,    Adjusted R-squared:  0.05533 
## F-statistic: 86.57 on 1 and 1460 DF,  p-value: < 2.2e-16
```

Interpretation

Since the p-value is less than the conventional threshold of 0.05, we can reject the null hypothesis that there is no correlation between the two variables. This means that there is statistically significant evidence to suggest that a relationship does exist between average rating and the number of ratings.

Linear Correlation

> We first look at the residual plot for this linear model to check whether using simple linear regression is appropriate. Logarithmic values of `Rating Count` are used for the following calculations and analysis.

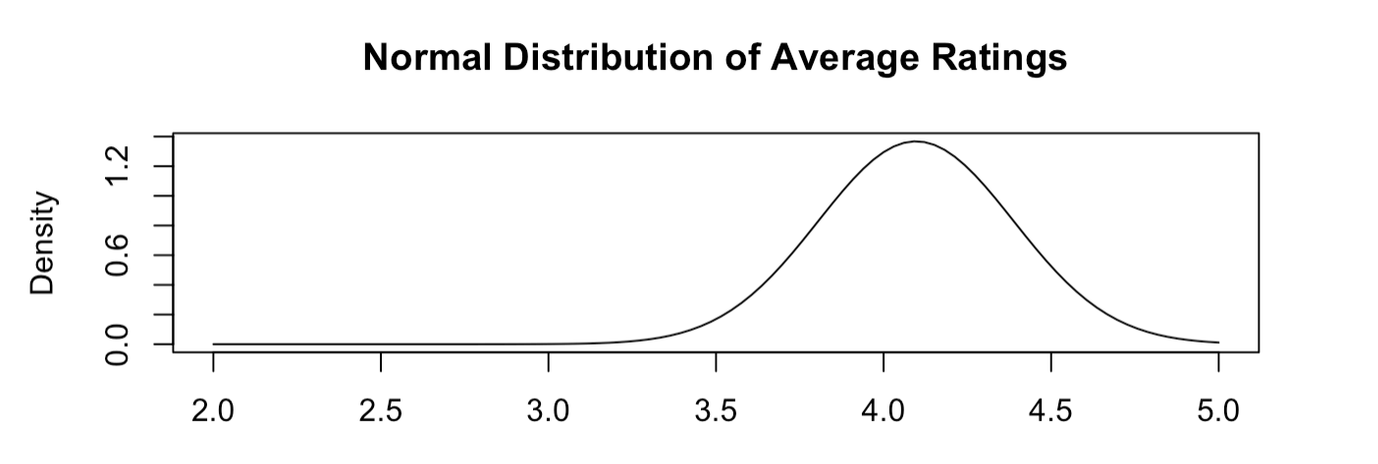


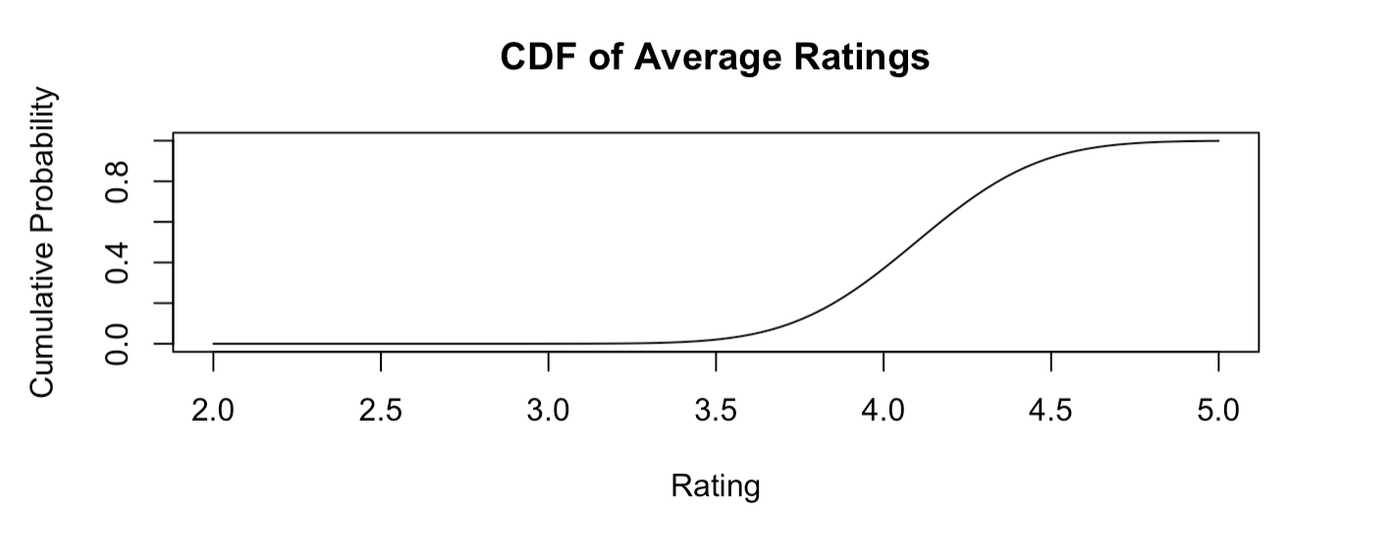
There is no obvious pattern in the residual plot, so using linear model should be appropriate. We go on to calculate the correlation coefficient using function `cor()`.

```
## [1] 0.2365939
```

The data from Amazon shows a slight positive linear correlation between Product Ratings and the Number of Ratings, with a correlation coefficient of 0.237. Although statistically significant, the relationship is weak, suggesting that while better ratings may slightly increase the number of ratings —which we're using to guess sales—they don't guarantee higher sales by themselves. This implies that other factors, like marketing and product visibility, also play significant roles in influencing sales.

> In our continued exploration of Amazon's product data, we turn our focus to understanding the likelihood of different rating levels. To visualize this, we'll construct two plots: one to show the probability distribution of ratings (how densely ratings are expected to occur around the average) and another to show the cumulative probability (the chance of a rating being at or below a certain level). These visual tools will help us grasp the chances of encountering products with certain ratings on Amazon, guiding potential strategies for sellers and insights for buyers.





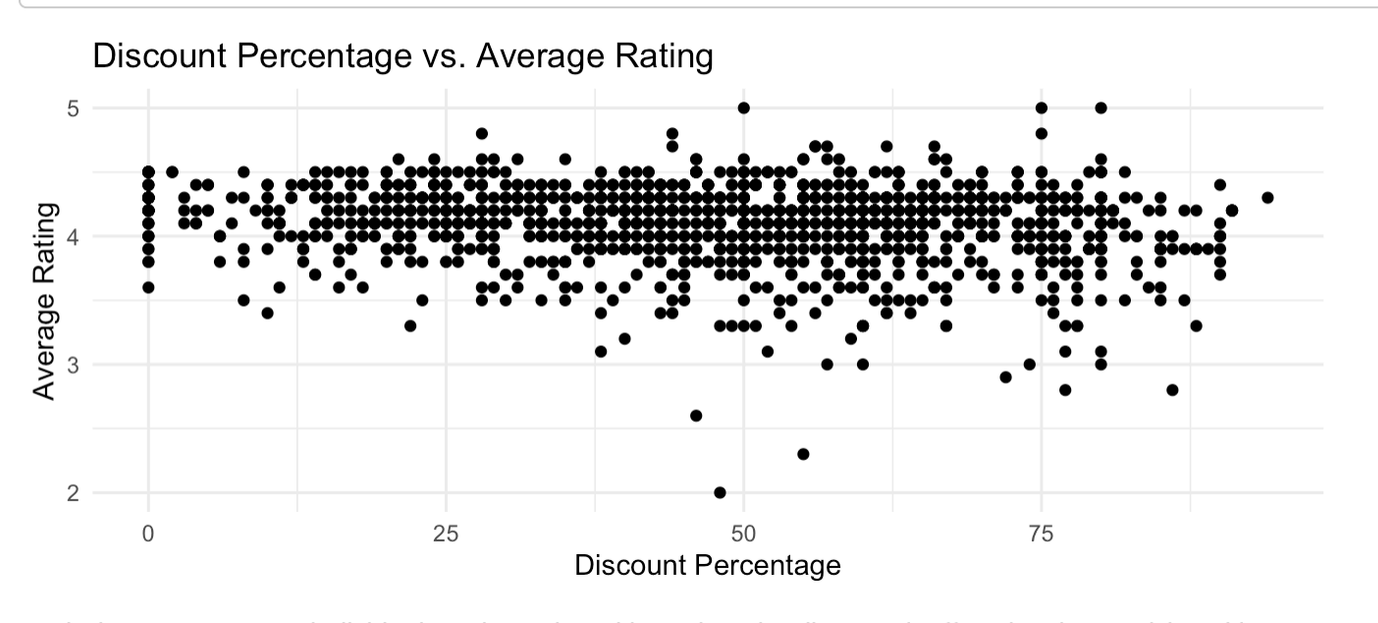
The "Normal Distribution of Average Ratings" plot suggests that most products have ratings around the middle of the scale, with fewer products at the very high or low ends. The "CDF of Average Ratings" plot shows the probability of a product having a rating up to a certain value, indicating most ratings are below 5. Together, these plots hint that on Amazon, products mostly receive moderate ratings, with extreme ratings being less typical.

Analysis - Question 2

> In the provided graphical summary, we have created a scatter plot to visualize the relationship between the discount percentage and the average customer rating of products on Amazon. This plot is designed to investigate if there's a discernible pattern or trend linking how much a product is discounted to how it is rated by customers. The scatter plot serves as a tool to explore complex relationships in the data, offering insights into how pricing strategies might affect customer perceptions and the overall reputation of products on Amazon.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   2.000   4.000   4.100   4.097   4.300   5.000       1
```

```
##    Length     Class      Mode 
##      1465 character character
```



Each dot represents an individual product, plotted based on the discount it offers (on the x-axis) and its average customer rating (on the y-axis).

From the plot, it appears that products are spread across a range of discount percentages without a clear trend indicating that discounts lead to higher ratings. Most ratings cluster around the 4 to 5-star mark, suggesting that customers generally leave positive reviews, regardless of the discount level. However, there is a visible presence of products with lower ratings that span across various discount levels, indicating that lower prices do not necessarily guarantee higher satisfaction.

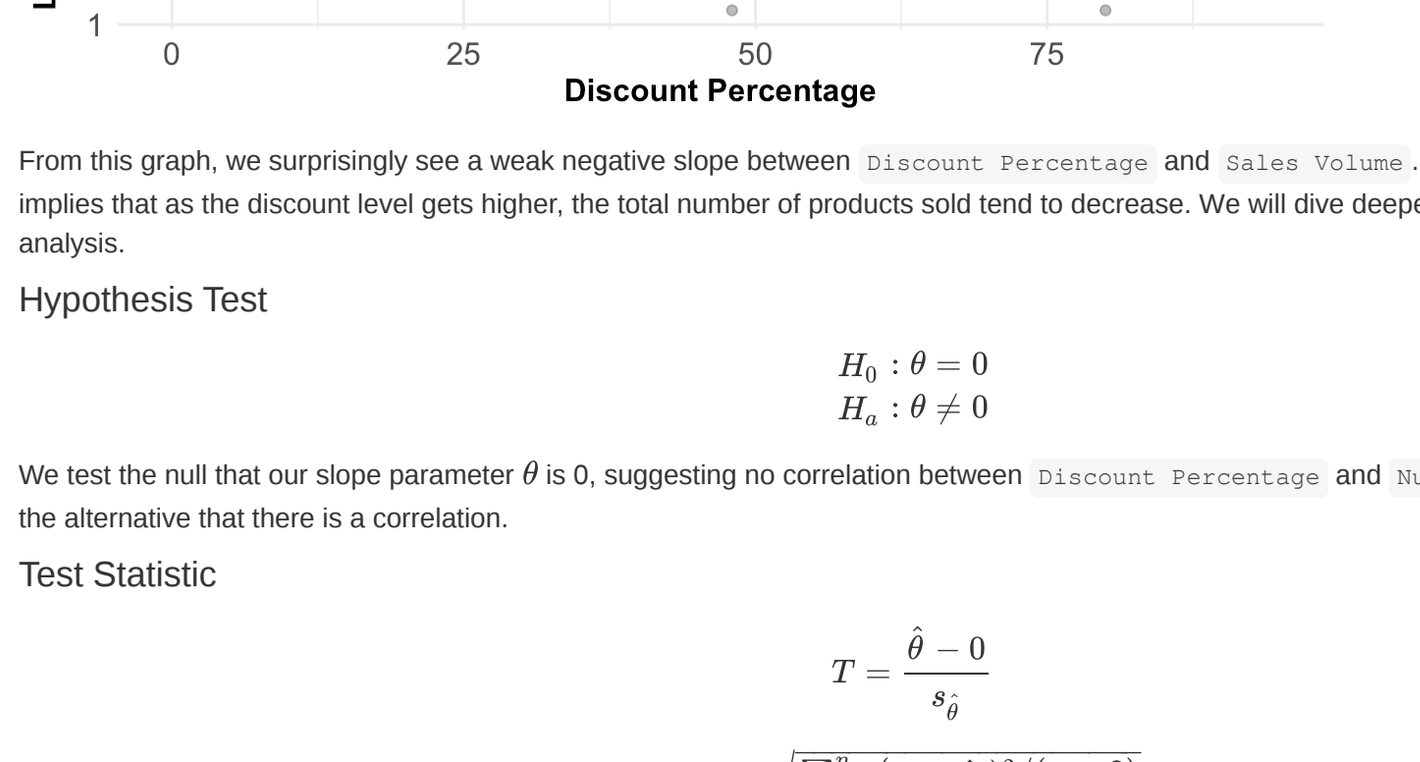
What's notable is the density of products with no or little discount that still maintain high average ratings, which might imply that factors other than price reductions contribute to customer satisfaction.

> Our initial scatter plot illustrated individual products' ratings in relation to the discounts offered, revealing a dispersed pattern without a clear trend. To distill this information further, we will now carry out rating count density. The volume of ratings per product can be a proxy for its popularity or sales volume. By examining the density of rating counts across different discount levels, we aim to identify whether there is a correlation between the number of ratings a product receives and the discount applied to it. This analysis might help us understand if discounts not only influence customer satisfaction. Again, we use logarithmic sclae for `Rating Count` to allow smaller values to be displayed clearly and to manage skewness caused by large values.

From this analysis, it appears that the level of discount offered on Amazon products does not have a significant impact on the number of ratings those products receive. This could suggest that other factors, such as product quality, brand reputation, or customer service, may play a more critical role in influencing customer engagement, as reflected by rating counts. This information can be particularly useful for sellers who might be considering discount strategies as a way to increase product visibility and customer engagement on platforms like Amazon.

> Next, we focus on the correlation between product discount percentage and their sales volume. We first create a scatter plots to show discount percentage agianst rating counts, which we use to represent sales volume.

On the next graph, the logarithmic scale would provide a better visualization of the correlation between `Discount Percentage` and `Sales Volume`. Logarithmic scale on the y-axis (rating count) allows smaller values to be displayed more clearly and helps to manage the skewness caused by large values.



From this graph, we surprisingly see a weak negative slope between `Discount Percentage` and `Sales Volume`. To be more specific, this implies that as the discount level gets higher, the total number of products sold tend to decrease. We will dive deeper into this in the following analysis.

Hypothesis Test

$$H_0: \theta = 0$$
$$H_a: \theta \neq 0$$

We test the null that our slope parameter θ is 0, suggesting no correlation between `Discount Percentage` and `Number of Ratings`, against the alternative that there is a correlation.

Test Statistic

$$T = \frac{\hat{\theta} - 0}{s_{\hat{\theta}}}$$

$$s_{\hat{\theta}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

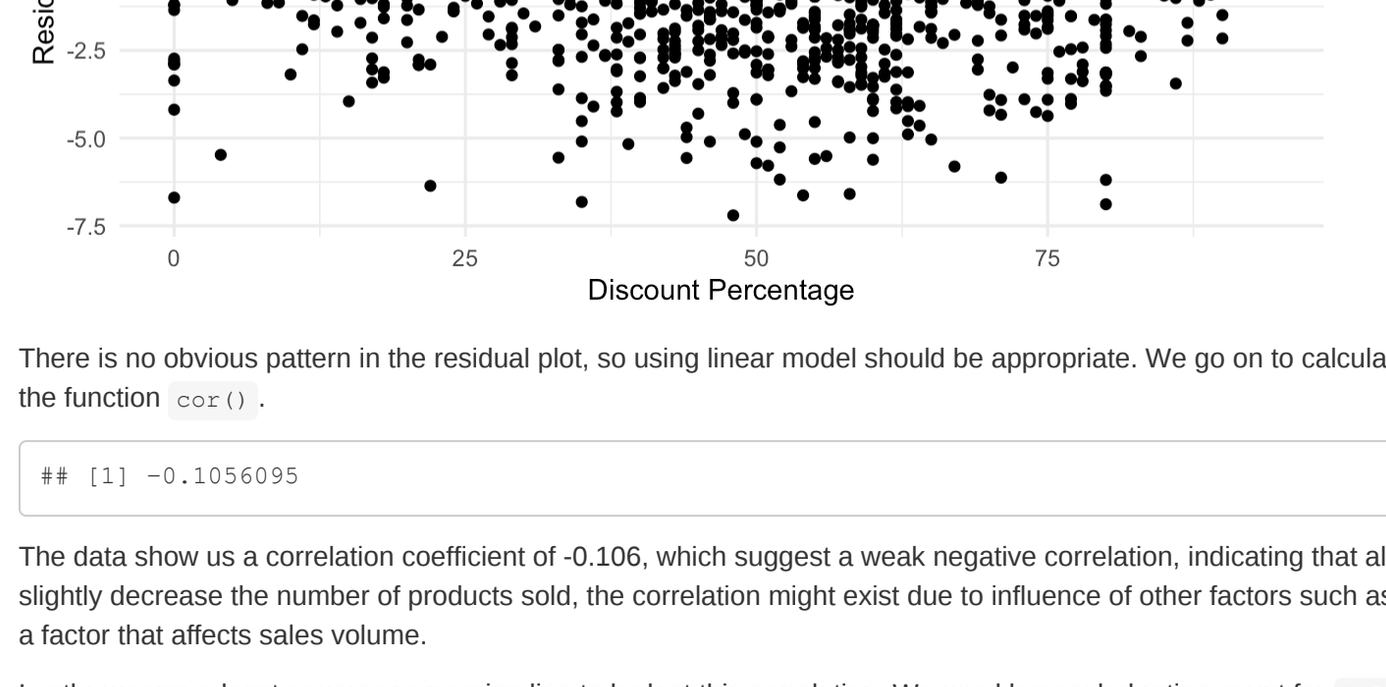
```
## 
## Call:
## lm(formula = log1p(rating_count) ~ discount_percentage, data = amazon)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -7.2008 -1.1753  0.2514  1.3908  4.9621 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          8.774588   0.127639  68.745  < 2e-16 ***
## discount_percentage -0.009900   0.002439  -4.058  5.2e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.017 on 1461 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.01115,    Adjusted R-squared:  0.01047 
## F-statistic: 16.47 on 1 and 1461 DF,  p-value: 5.202e-05
```

Interpretation

Since the p-value is less than the conventional threshold of 0.05, we can reject the null hypothesis that there is no correlation between the two variables. This means that there is statistically significant evidence to suggest that a relationship does exist between discount percentage and the number of ratings (sales volume).

Linear Correlation

> We first look at the residual plot for this linear model to check whether using simple linear regression is appropriate. Logarithmic values of `Rating Count` are used for the following calculations and analysis.



There is no obvious pattern in the residual plot, so using linear model should be appropriate. We go on to calculate the correlation coefficient using the function `cor()`.

```
## [1] -0.1056095
```

The data show us a correlation coefficient of -0.106, which suggest a weak negative correlation, indicating that although higher discount levels may slightly decrease the number of products sold, the correlation might exist due to influence of other factors such as product quality, which would be a factor that affects sales volume.

Lastly, we use a least-squares regression line to look at this correlation. We used log-scaled rating count for `Sales Volume` to stay consistent with the scatter plot.

Discussion

Interpretation of Results

The analysis of the Amazon sales data focused on a range of products, primarily in the electronics and accessories category. The study highlighted key aspects such as product popularity, pricing strategies, and customer reviews. The results should be interpreted in the context of consumer preferences and market trends specific to the e-commerce sector. For instance, the high demand for certain products could be attributed to their features, price points, or customer satisfaction levels as indicated by reviews and ratings. As for the application of this Amazon Sales Analysis, the results we have obtained can be most directly applied in the business world. For example, the correlation or trends provided in this analysis can help manufacturers adjust sales plans or influence stakeholders on their decision-making.

Potential Shortcomings

- The primary limitation of our analysis is the reliance on Amazon ratings as the sole metric of customer satisfaction and engagement. This is because customer ratings are subjective, which is a source for bias as they can be influenced by numerous factors beyond the product's inherent quality. For instance, customers more inclined to leave reviews may have extreme opinions, potentially skewing the average rating data towards a more positive or negative end. Additionally, the analysis does not account for other factors like worded customer reviews, return rates, and repeat purchases, which could provide a more holistic view of customer satisfaction.
- Another shortcoming is the restricted scope of the data set. Focusing primarily on specific categories may not provide a comprehensive view of Amazon's diverse product range.
- Lastly, the temporal scope of the data is another constraint, as it may not capture long-term trends or seasonal variations effectively.

Future Directions

- New Questions: Future research could explore different product categories or compare online sales trends with in-store sales data. Investigating the impact of promotional events or changes in market conditions (like economic downturns) on sales would also be insightful.
- Different Methods: Utilizing advanced data analytics techniques, such as machine learning models for predictive analysis or sentiment analysis on customer reviews, could provide deeper insights. Integrating data visualization tools for more interactive and intuitive representation of data could also be beneficial.
- New Data: Collecting data over a more extended period or incorporating data from other e-commerce platforms could enhance the analysis. Gathering demographic information about customers could also provide valuable insights into different market segments.

Conclusion

The primary conclusion from the analysis is the identification of key factors influencing product popularity and sales on Amazon. Products with competitive pricing, positive customer reviews, and high ratings tend to perform better in sales. The data also suggests that customer feedback, both in terms of ratings and written reviews, plays a critical role in influencing potential buyers' decisions. These conclusions are primarily supported by the analysis of product ratings, review content, and pricing strategies observed in the dataset. We believe that it's important to consider these findings within the broader context of the rapidly evolving e-commerce landscape and the limitations of the dataset.