EXPLORATORY DATA ANALYSIS: AMAZON SALES DATASET

OUTLINE

1. **Dataset Overview:**

1.1 Summary statistics of numerical features (mean, median, min, max, etc.).

1.2 Data cleaning and manipulation

1.3 Data types of each feature.

2 **Univariate Analysis:**

2.1 Distribution of sentiment.

2.2 Exploration of the top ten categories of products available in the dataset.

2.3 Distribution of Ratings.

2.4 Exploration of the distribution of the discounted prices.

3 **Bivariate Analysis:**

3.1 Correlation analysis between numerical features

3.2 Relationship between ratings and other numerical features (e.g., price).

3.3 Compare the average ratings across different product categories.

3.4 Sentiment analysis of name and score

4 **Text Analysis:**

4.1 Explore the length of review titles and review contents.

4.2 Most common words used in review titles and contents.

4.3 Sentiment analysis of reviews (positive, negative, neutral).

5 **Feature Engineering:**

5.1 User engagement

5.2 User activity level

6 **Insights and Conclusions:**

6.1 Key insights

6.2 Recommendation

7 **Conclusion:**

7.1 Recap of the EDA process and its findings.

7.2 Highlight the significance of the analysis and potential next steps.

1. Dataset Overview:

1.1 Summary statistics of numerical features (mean, median, min, max, etc.).

Discounted price	Actual price	Discount percentage	rating	Rating count	Discounted price
mean	3078.073	5359.106	0.476532	4.094186	11949.67
median	799	1670	0.5	4.1	4744
min	39	39	0	1	2
max	61999	85000	0.94	5	98250

i. **Discounted Price:**

- The mean discounted price of products in the dataset is approximately ₹3,078.07, with a median value of ₹799. This suggests that there is considerable variability in the prices of discounted products, with some being significantly lower than the median value.
- The minimum discounted price recorded is ₹39, indicating that there are products available at very low prices, possibly due to heavy discounts or promotional offers.
- On the higher end, the maximum discounted price observed is ₹61,999, indicating that there are also products available at relatively high prices, possibly due to their premium quality or brand value.

ii. **Actual Price:**

- The mean actual price of products in the dataset is approximately ₹5,359.11, with a median value of ₹1,670. This suggests that there is a notable difference between the discounted and actual prices of products.
- The minimum actual price recorded is ₹39, which is consistent with the minimum discounted price, indicating that there are products available at the same minimum price regardless of discounts.
- On the higher end, the maximum actual price observed is ₹85,000, indicating that there are also products available at significantly higher prices without discounts.

iii. **Discount Percentage:**

- The mean discount percentage for products in the dataset is approximately 47.65%, indicating that, on average, products are discounted by nearly half of their original prices.
- The median discount percentage is 50%, suggesting that there is an equal distribution of discounts above and below this value.

- The minimum discount percentage recorded is 0%, indicating that some products may not be discounted at all, while the maximum discount percentage is 94%, indicating substantial discounts on certain products.

iv. **Rating:**

- The mean rating for products in the dataset is approximately 4.09 out of 5, indicating generally positive feedback from customers.
- The median rating is also 4.1, suggesting that the distribution of ratings is relatively symmetric around this value.
- The minimum rating recorded is 1, indicating the lowest possible rating, while the maximum rating is 5, indicating the highest possible rating.

v. **Rating Count:**

- The mean number of ratings received by products in the dataset is approximately 11,949.67, indicating a significant level of customer engagement and feedback.
- The median rating count is 4,744, suggesting that there is a wide range in the number of ratings received by products, with some having substantially more ratings than others.
- The minimum rating count recorded is 2, indicating that there are products with very few ratings, while the maximum rating count is 98,250, indicating a product with exceptionally high engagement and feedback.

1.2 Data cleaning and manipulation

During the data cleaning process, several issues were identified and addressed:

i. It was discovered that the letter '1' in ratings was represented by '|' character. This inconsistency was rectified by replacing all occurrences of '|' with '1', ensuring accurate representation of ratings.

ii. The proper currency symbol ('₹') was not encoded and displayed as 'â,¹'. Attempts to change the currency symbol failed, necessitating an alternative approach. The solution involved using text splitting techniques to separate the values from the currency symbol. Despite efforts to insert the correct currency symbol afterward, persistent issues led to the decision to remove the currency symbol altogether.

iii. As certain columns, such as 'product_link' and 'img_link', were deemed irrelevant to the analysis, they were removed from the dataset to streamline further processing and analysis.

iv. Additionally, the presence of two blank entries in the 'rating' column posed data integrity concerns. To maintain the dataset's integrity, these rows were promptly removed to ensure accurate and reliable analysis.

These measures were undertaken to enhance the quality and usability of the dataset, facilitating more accurate analysis and interpretation of the data.

1.3 Data types of each feature.

i. **Product Information:**

- **product_id**: Unique identifier for each product (string).
- **product_name**: Name of the product (string).
- **category**: Category of the product (string).
- **discounted_price**: Discounted price of the product (numeric decimal).
- **actual_price**: Actual price of the product (numeric decimal).
- **discount_percentage**: Percentage of discount for the product (numeric decimal).
- **about_product**: Description about the product (string).
- **img_link**: Image link of the product (string).
- **product_link**: Official website link of the product (string).

ii. **Review Information:**

- **user_id**: ID of the user who wrote the review for the product (string).
- **user_name**: Name of the user who wrote the review for the product (string).
- **review_id**: ID of the user review (string).
- **review_title**: Short review title (string).
- **review_content**: Long review content (string).
- **rating**: Rating of the product (string).
- **rating_count**: Number of people who voted for the Amazon rating (numeric whole).

iii. Derived columns

- **Sentiment (string):**
  The sentiment column indicates the overall sentiment expressed in the review. It categorizes the sentiment into different categories such as positive, negative, or neutral based on the sentiment score.

- **Sentiment Name (string):**
  The sentiment name column provides descriptive labels for the sentiment categories. In this case, for positive sentiment, common labels include "good," "excellent," "nice," etc. These labels help to interpret and understand the sentiment associated with each review.

- **Sentiment Score (number whole):**
  The sentiment score column quantifies the number of times the sentiment name was expressed in the dataset. It represents the frequency or occurrence of specific positive (or negative) words used in the review content.

- **Specific Category(string):**
  The specific category column was derived from the original category column through a text parsing or text splitting process. The original category column contained hierarchical information about the product categories, with multiple levels of categorization separated by "|" (pipe) characters. Each level represented a different hierarchy or subcategory of the product.
  To derive the specific category, the text within the category column was split or parsed based on the "|" delimiter. This splitting process separated the text into individual components, allowing for the extraction of the most granular level of categorization, which corresponds to the specific category.

- **User engagement (string)**
  The user engagement column categorizes average rating counts into low, medium, and high interaction levels for analysis.

- **Review title length (number decimal)**
  The length of the review title indicates the number of characters used in the title.

- **Review content length (number whole)**
  Review content length refers to the character count within the body of the review.

- **Original price (number decimal)**
  Original price is calculated by dividing the discounted price by (1 - discount percentage / 100).

- **Revenue per unit (number decimal)**
  Revenue per unit is computed by dividing the original price by (1 - discount percentage / 100).

- **Discount amount:**

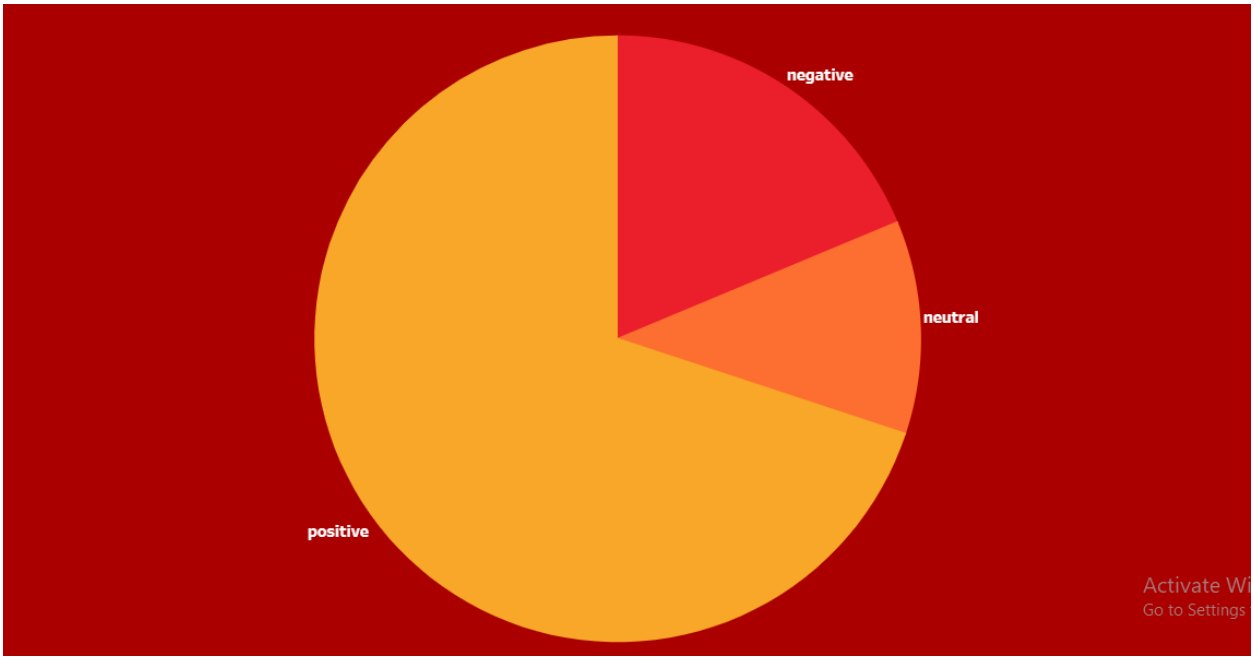
The discount amount is determined by subtracting the discounted price from the actual price.

- **Corr review title and content (number decimal):**

The correlation between review title and content length measures the relationship between these two variables, providing insights into user behavior or preferences.

2. Univariate Analysis:

2.1 Distribution of Sentiment



The provided data summarizes the sentiment expressed in reviews along with their corresponding sentiment scores. Here's an expanded explanation:

- The data indicates that a significant number of reviews express a positive sentiment towards the product or experience, with a sentiment score of 1,342.
- There are fewer reviews categorized as negative, with a sentiment score of 359, indicating a lower intensity of negativity compared to the positivity expressed in the positive category.
- The neutral category has a sentiment score of 218, suggesting that a moderate number of reviews express a neutral sentiment without strong emotional tone.

➢ **Implications**:

- The distribution of sentiment scores provides insights into the overall sentiment of the reviews and customer perceptions of the product or experience.
- A high number of positive sentiment scores may indicate customer satisfaction and positive experiences with the product, potentially contributing to brand loyalty and positive word-of-mouth.
- Negative sentiment scores, although lower in number, may highlight areas for improvement or dissatisfaction that require attention from the company to address customer concerns and enhance the product or experience.

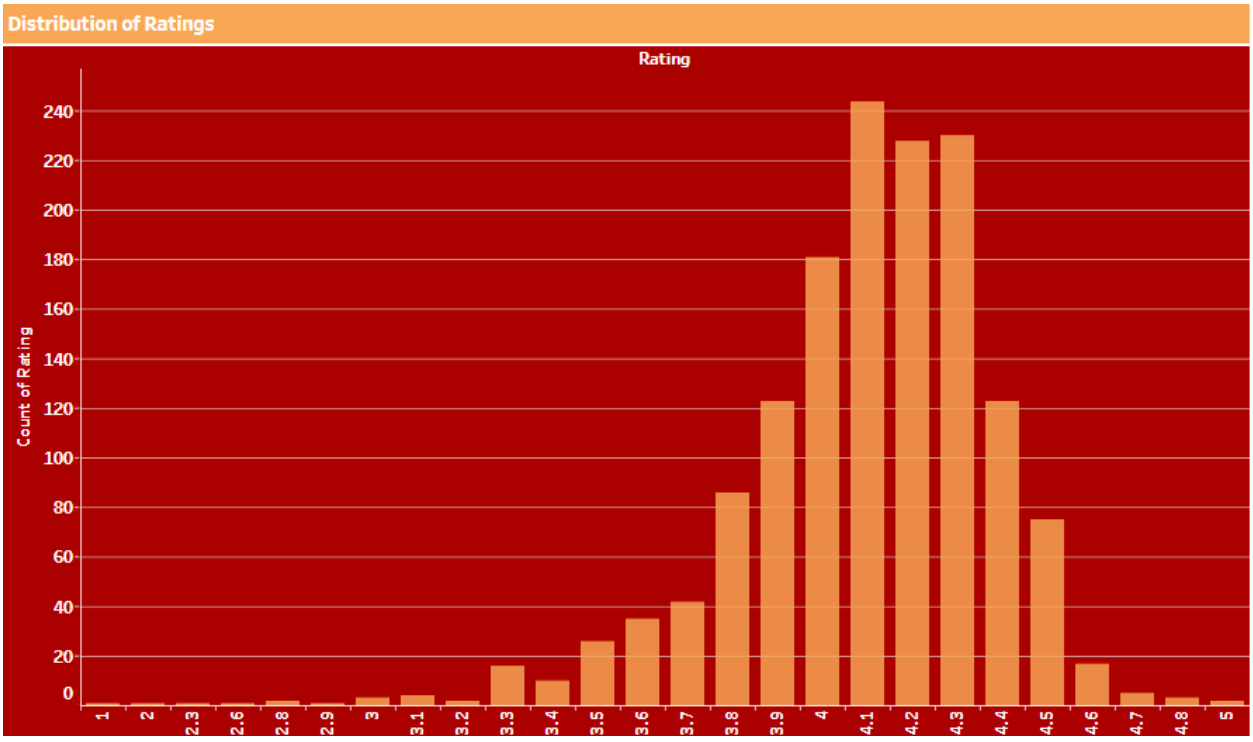
2.2 Exploration of the top ten categories of products available in the dataset.

The provided data presents the top ten categories along with their respective counts. Let's expand on this information:

- Home & Kitchen | Kitchen & Home Appliances | Vacuum, Cleaning & Ironing | Irons, Steamers & Accessories | Irons | Dry Irons (Count: 24):
- Home & Kitchen | Kitchen & Home Appliances | Small Kitchen Appliances | Mixer Grinders (Count: 27)
- Electronics | Wearable Technology | Smart Watches (Count: 76)
- Electronics | Mobiles & Accessories | Smartphones & Basic Mobiles | Smartphones (Count: 68)
- Electronics | Home Theater, TV & Video | Televisions | Smart Televisions (Count: 62)
- Electronics | Home Theater, TV & Video | Accessories | Remote Controls (Count: 49)
- Electronics | Home Theater, TV & Video | Accessories | Cables | HDMI Cables (Count: 24)
- Electronics | Headphones, Earbuds & Accessories | Headphones | In-Ear (Count: 52)
- Computers & Accessories | Accessories & Peripherals | Keyboards, Mice & Input Devices | Mice (Count: 24)
- Computers & Accessories | Accessories & Peripherals | Cables & Accessories | Cables | USB Cables (Count: 231)

This category encompasses USB cables, which are used for connecting electronic devices, such as computers, smartphones, and peripherals, for data transfer or charging purposes.These top ten categories provide insights into the diverse range of products available across different segments, including kitchen appliances, electronics, wearable technology, and computer accessories.

2.3 Distribution of ratings.



i. The summary of rating provided represents the distribution of ratings across different values, along with the corresponding number of reviews for each rating. This data was visualized using a bar chart, which effectively illustrates the frequency of ratings and helps identify any patterns or trends in customer feedback.

Key observations from the summary of rating:

- **Ratings Distribution:** The ratings range from 1 to 5, indicating the full spectrum of customer feedback on the products. This range encompasses both positive and negative reviews, providing a comprehensive view of customer sentiment.

- **Majority of Ratings:** The majority of ratings fall within the range of 4 to 5, indicating that most customers have provided positive feedback for the products. This suggests a high level of satisfaction among customers overall.

- **Variation in Rating Counts:** There is significant variation in the number of reviews for each rating. While some ratings, such as 4.4 and 4.3, have millions of reviews, others, such as 3.1 and 2.6, have relatively fewer reviews. This variation reflects differences in product popularity, customer preferences, and the distribution of ratings among different products.

- **Distribution Shape:** The distribution of rating counts appears to follow a skewed pattern, with a larger number of reviews concentrated towards higher ratings (4 and above) and fewer reviews towards lower ratings (below 4). This suggests that the products are generally well-received by customers, with fewer instances of negative feedback.

2.4 Exploration of the distribution of the discounted prices.

- **Price Distribution:** The discounted prices vary widely, ranging from as low as $39 to as high as $61,999. This indicates a diverse range of products or services being offered.

- **Variability in Discounts:** Some items have substantial discounts, such as those with discounted prices significantly lower than their original prices, while others have minimal or no discounts at all.

- **Common Price Points:** Certain price points appear frequently in the dataset, such as $99, $199, $299, $499, $999, $1,999, and $9,999. These may represent common pricing strategies or product categories.

- **Occasional High Discounts:** There are instances where the discounted price is significantly lower than the original price, suggesting promotional sales or clearance events.

- **High-Value Items:** Several items have extremely high discounted prices, such as $61,999 and $103,992. These could be luxury products or high-end services.

- **Clustering of Prices:** Prices seem to cluster around certain values, such as $99, $199, $299, $999, etc., indicating pricing strategies based on psychological pricing or market segmentation.

- **Potential Revenue Impact:** Understanding the distribution of discounted prices can provide insights into potential revenue impact, especially if higher discounts lead to increased sales volume.

- **Market Competition:** The presence of items with high discounts could indicate intense competition in certain markets, where businesses are using discounts as a strategy to attract customers.

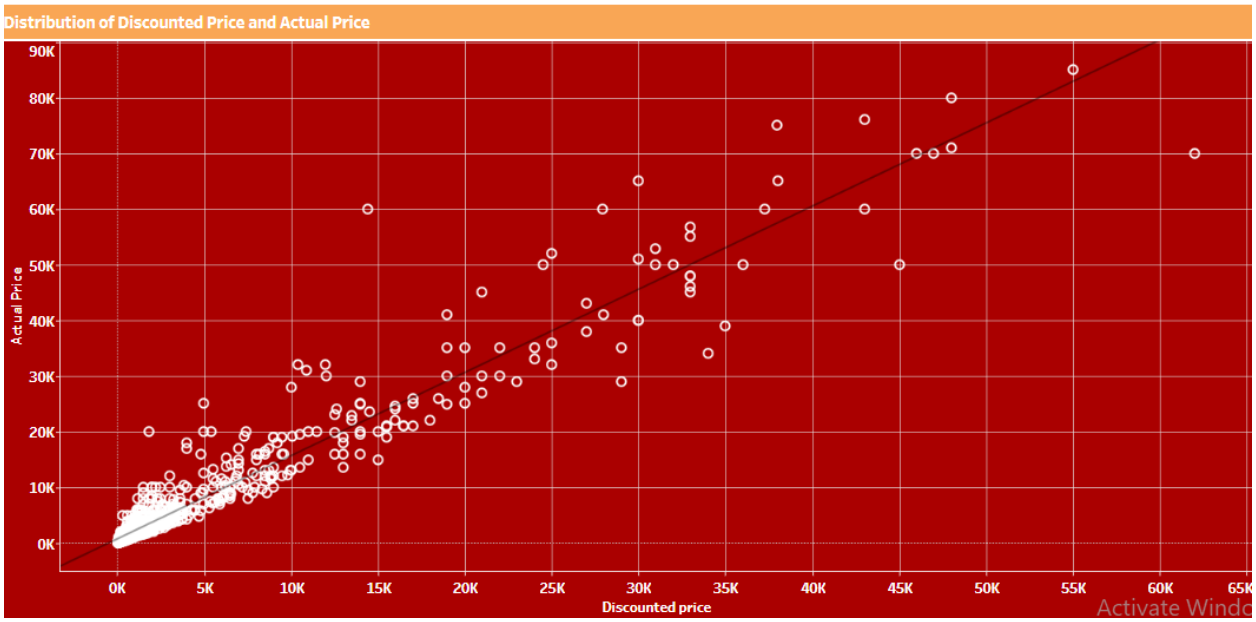
- **Segmentation Opportunities:** Analyzing the data further by product categories or customer segments may reveal patterns in pricing strategies and help identify opportunities for targeted marketing or pricing adjustments.

- **Customer Perception:** Perception of value can vary based on the discount offered. Understanding how customers perceive different levels of discounts can inform pricing strategies and marketing messaging.

3. Bivariate variables

3.1 Correlation analysis between numerical features

i. Actual price and discounted price



a) **Consistent Distribution Within Ranges:**

The majority of data points exhibit a consistent distribution within specific ranges, particularly between 0k to 10k for discounted prices and 0k to 15k for actual prices. This consistency indicates that products within these price ranges are commonly encountered in the dataset, reflecting a cohesive distribution of pricing across the plot.

b) **Strong Positive Correlation:**

The clustering pattern observed along the trend line reflects a strong positive correlation between discounted prices, actual prices, and potentially other variables such as product categories or features. As discounted prices increase, there is a corresponding increase in actual prices, indicating a consistent pricing strategy or trend among the products analyzed.

c) **Cohesive Distribution of Outliers:**

Even the outliers, though not tightly clustered, do not deviate significantly from one another. This cohesive distribution suggests that while there may be variations in pricing or other factors for certain products, these outliers still align with the overall trend observed in the plot. It indicates a degree of consistency in product pricing and characteristics across the dataset.

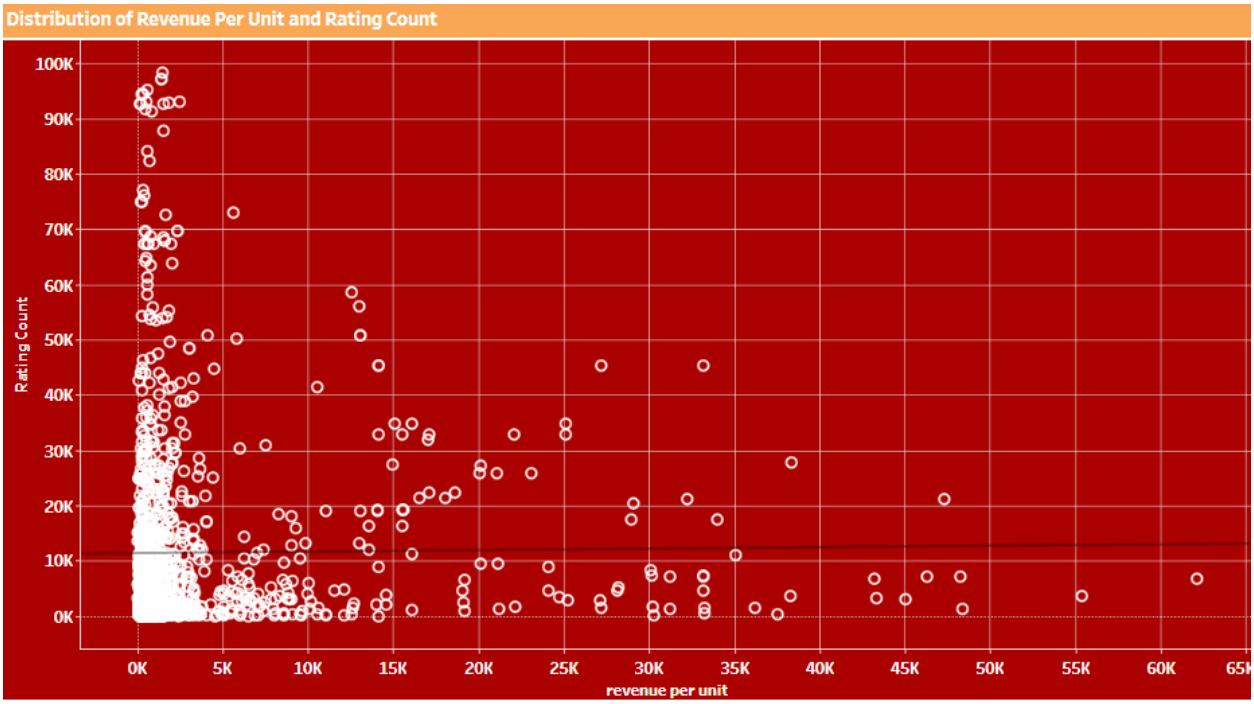
d) **Implications for Market Dynamics:**

The clustering pattern and strong positive correlation imply a level of cohesion and consistency within the market segment represented by the dataset. This may reflect common pricing strategies, consumer preferences, or market dynamics that influence the pricing and distribution of products within certain price ranges.

e) **Potential Factors Driving Correlation:**

The observed correlation between discounted prices, actual prices, and other variables may be influenced by various factors such as product features, brand reputation, competitive pricing strategies, and market demand. Understanding these factors can provide insights into market trends, pricing dynamics, and consumer behavior within the industry.

3.2 Relationship between ratings and other numerical features (e.g., price).



i. Distribution of discount percentages to rating.

The provided summary displays the discount percentage associated with each rating level. This data was visualized using a bar chart, which effectively communicates the relationship between ratings and their corresponding discount percentages.

Key observations from the summary of discount percentages by rating:

a) **Variation in Discount Percentage:** There is significant variation in the discount percentages across different rating levels. Higher ratings tend to have lower discount percentages, while lower ratings often correspond to higher discount percentages. This suggests that products with lower ratings may be subject to deeper discounts as a strategy to incentivize purchases or mitigate negative feedback.

b) **Inverse Relationship:** Overall, there appears to be an inverse relationship between ratings and discount percentages. As the rating decreases, the discount percentage tends to increase, and vice versa. This observation aligns with common marketing strategies aimed at boosting sales by offering discounts on products that may have received less favorable reviews.

c) **Threshold Effect:** Certain rating levels, such as 4.5 and above, exhibit relatively low discount percentages, indicating that these products may be highly regarded by customers and therefore less likely to require significant discounts to drive sales. Conversely, lower rating levels, such as 3.5 and below, show higher discount percentages, suggesting that these products may face challenges in attracting customers without price incentives.

d) **Impact on Purchase Behavior**: The relationship between ratings and discount percentages can influence customer purchase behavior. Products with higher ratings and lower discount percentages may be perceived as higher quality and therefore command premium prices. Conversely, products with lower ratings and higher discount percentages may be viewed as offering less value for money, potentially impacting consumer trust and brand loyalty.

Overall, the bar chart visualization provides a clear representation of the relationship between ratings and discount percentages, offering insights into pricing strategies, customer perceptions, and potential areas for improvement in product offerings.

ii. Revenue per minute and rating count

Upon examining the scatter plot and trend line, several key observations were made regarding the relationship between rating count and revenue per unit:

- **Straight Trend Line:**

The trend line displayed a linear relationship between rating count and revenue per unit. This indicates that as the rating count increases, there is a corresponding increase in revenue per unit, following a consistent pattern or trend.

- **Concentration of Data:**

A concentration of data points was observed within the range of 20,000 to 40,000 rating count and 0 to 5,000 revenue per unit. This concentration suggests that a significant portion of the observations falls within this range, indicating a commonality or trend among these data points.

- **Data Distribution at Lower Revenue Levels:**

The data points were tightly packed together, particularly within the lower revenue range (0 to 10,000 per unit). This clustering of data indicates that products with lower revenue per unit values tend to have similar rating counts, suggesting a consistent level of customer engagement or feedback across these products.

- **Scattered Data at Higher Revenue Levels:**

As the revenue per unit exceeded 10,000, the data points became more scattered, indicating greater variability in rating counts for products with higher revenue levels. This dispersion suggests that products generating higher revenue per unit may exhibit more diverse customer ratings, with some products receiving significantly higher or lower ratings compared to others in the dataset.

- **Interpretation:**

The concentration of data within certain ranges and the dispersion observed at higher revenue levels can be interpreted in several ways. It may suggest that products with moderate revenue per unit values experience a consistent level of customer engagement, resulting in a clustered distribution of rating counts.

Conversely, the increased variability in rating counts at higher revenue levels may indicate differences in product quality, customer satisfaction, or market demand among higher-priced products, leading to a wider dispersion of data points.

iii. Rating Categories and Discounted Prices

a) **Water Purifiers & Accessories:**

- The majority of products in this category have ratings ranging from 2.9 to 4.6, with corresponding discounted prices varying widely from ₹199.00 to ₹26,402.00.
- There is a notable variation in discounted prices within this category, indicating diverse product offerings or features that may influence pricing.

b) **Water Heaters & Geysers:**

- Products in this category exhibit ratings predominantly between 3.6 and 4.8, with discounted prices spanning from ₹600.00 to ₹36,558.00.
- Higher-priced items, such as those priced above ₹10,000.00, tend to have ratings towards the higher end of the spectrum, suggesting a potential correlation between price and perceived quality.

c) **Vacuum, Cleaning & Ironing:**

- Ratings for vacuum, cleaning, and ironing products vary between 3.1 and 4.6, with discounted prices ranging from ₹199.00 to ₹38,687.00.
- There is a wide range of prices within this category, possibly reflecting differences in product types (e.g., vacuum cleaners, irons) and features (e.g., cordless, robotic).

d) **Televisions:**

- Televisions display a broad range of ratings, from 3.4 to 4.7, with corresponding discounted prices spanning from ₹7,299.00 to ₹764,845.00.
- High-end televisions with prices exceeding ₹100,000.00 tend to have ratings towards the higher end of the spectrum, suggesting a correlation between price point and perceived quality or features.

e) **Small Kitchen Appliances:**

- Products in this category have ratings ranging from 2.3 to 4.8, with discounted prices varying widely from ₹161.00 to ₹69,845.00.
- There is a considerable variation in discounted prices within this category, reflecting the diverse range of kitchen appliances available, from blenders to coffee makers.

f) **Mobile Accessories:**

- Ratings for mobile accessories span from 3.6 to 4.7, with discounted prices ranging from ₹199.00 to ₹236,898.00.
- Higher-priced mobile accessories tend to have ratings towards the higher end of the spectrum, suggesting a correlation between price point and perceived quality or functionality.

iv. Rating Distribution by Specific Category

Here's an expanded explanation of the specific categories:

➢ **Highest Counts:**

a) **Cables & Accessories (Count: 238):**

- This category encompasses various types of cables and accessories, such as USB cables, HDMI cables, adapters, and connectors.

b) **Small Kitchen Appliances (Count: 181):**

- Small kitchen appliances include a wide range of devices used for food preparation and cooking, such as blenders, toasters, coffee makers, and electric kettles.

c) **Smartphones & Basic Mobiles (Count: 77):**

- This category includes smartphones and basic mobile phones, along with accessories such as cases, screen protectors, and chargers.

d) **Smart Watches (Count: 76):**

- Smartwatches are wearable devices that offer features beyond traditional timekeeping, such as fitness tracking, notifications, and app integration.

e) **Mobile Accessories (Count: 84):**

- Mobile accessories include a variety of products designed to enhance the functionality or protection of mobile devices, such as cases, chargers, cables, and screen protectors.

➢ **Lowest Counts:**

a) **Health Monitors (Count: 1):**

- Health monitors are devices used to track various health metrics, such as blood pressure monitors, heart rate monitors, and pulse oximeters.

b) **Flashes (Count: 1):**

- Flashes are photography accessories used to provide additional illumination in low-light situations, typically mounted on cameras or used off-camera.

c) **Drawing & Painting Supplies (Count: 1):**

- This category includes various art supplies used for drawing and painting, such as pencils, brushes, canvases, and paints.

d) **Data Cards & Dongles (Count: 1):**

- Data cards and dongles are portable devices that provide internet connectivity through cellular networks, commonly used for mobile internet access.

e) **Air Conditioners (Count: 1):**

- Air conditioners are appliances used for cooling indoor spaces, commonly found in homes, offices, and commercial buildings.

These counts provide insights into the popularity or prevalence of specific categories within the dataset, with some categories being more common or widely represented than others.

v. Rating vs. Discount Percentile Distribution

a. **Inverse Relationship:**

- As the discount percentage increases, there is a trend of ratings decreasing. This suggests an inverse relationship between discounts offered and customer satisfaction, indicating that products with higher discounts may receive lower ratings.

b. **Impact of Discounts on Ratings:**

- Products with lower discount percentages (e.g., 0.160% - 2.970%) tend to have higher ratings, indicating that customers perceive these products more positively despite lower discounts.
- Conversely, products with higher discount percentages (e.g., 111.450% - 111.980%) tend to have lower ratings, suggesting that customers may perceive these products as having lower value or quality despite the higher discounts.

c. **Threshold Effect:**

- There appears to be a threshold effect around certain discount percentage ranges (e.g., 30.780% - 111.980%), where ratings sharply decrease as discount percentages exceed a certain threshold. This could indicate that excessively high discounts may negatively impact customer perceptions of product quality or value.

d. **Distribution of Ratings:**

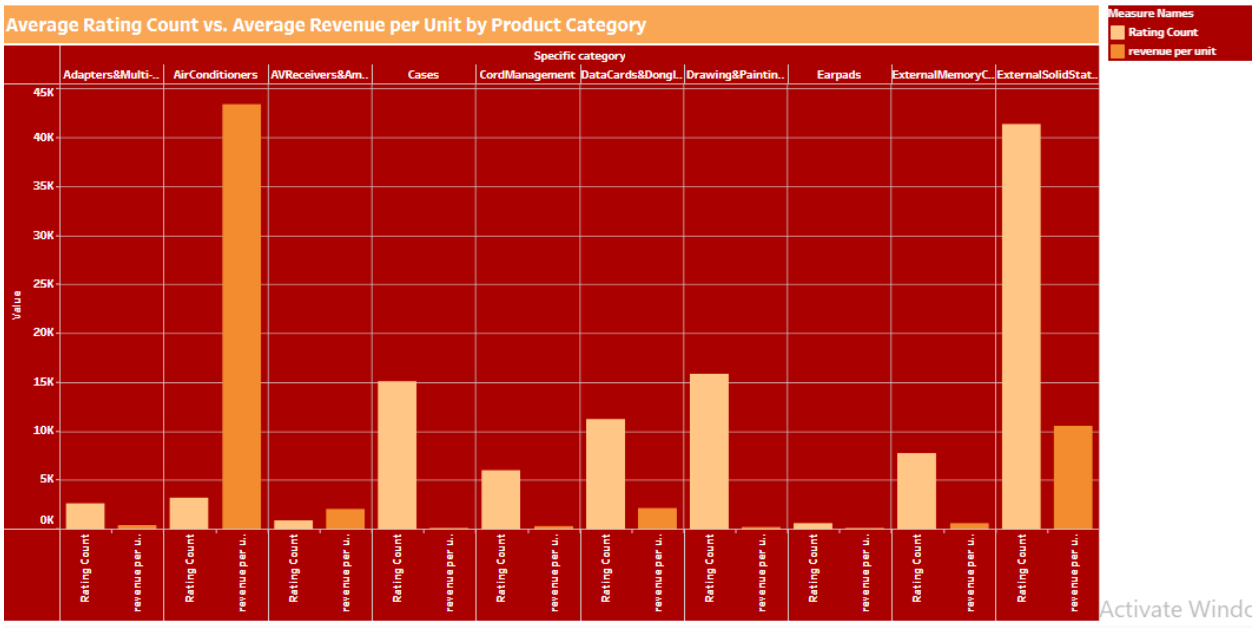
- The distribution of ratings varies across different discount percentage ranges, with some ranges having a wider spread of ratings than others. This suggests that the impact of discounts on ratings may vary depending on the specific context or product characteristics.

Implications:

- Understanding the relationship between ratings and discount percentages is crucial for pricing and promotional strategies. While discounts can attract customers, excessive discounting may erode perceived product value and lead to lower ratings.
- Businesses should carefully consider the balance between offering competitive discounts and maintaining product quality and value perception to maximize customer satisfaction and loyalty.

3.3 Compare the average ratings and revenue across different product categories.

I) Revenue Performance by Category:



The highest average revenue per unit is observed in the categories of Smartphones & Basic Mobiles, Televisions, and Small Kitchen Appliances. This indicates that these categories are potentially high-profit segments for Amazon.

Conversely, categories like Cables & Accessories and Accessories have relatively lower average revenue per unit, suggesting either lower-priced items or potentially lower demand compared to other categories.

ii. Customer Engagement and Satisfaction:

Categories such as Headphones, Smartwatches, and Smartphones & Basic Mobiles exhibit high average rating counts, indicating significant customer engagement and potentially high satisfaction levels with these products.

On the other hand, categories like Accessories and Cables & Accessories have lower average rating counts, which could imply lower customer engagement or satisfaction levels. Further investigation may be needed to understand the reasons behind this trend.

iii. Market Trends and Preferences:

The data suggests that certain categories, such as Smartphones & Basic Mobiles and Smartwatches, command higher prices (as indicated by their high average revenue per unit) and also attract a substantial number of ratings. This could indicate strong market demand for premium electronic devices.

Small Kitchen Appliances also stand out with high average revenue per unit, suggesting that consumers are willing to invest in quality kitchen appliances.

iv. Potential Growth Opportunities:

Categories with relatively lower average revenue per unit but high average rating counts, such as Accessories and Mobile Accessories, may present growth opportunities. Amazon could explore strategies to increase revenue per unit in these categories while maintaining or improving customer satisfaction levels.

Additionally, categories with moderate average revenue per unit but low average rating counts, like Vacuum, Cleaning & Ironing, may require attention to improve customer satisfaction and engagement, potentially through product quality enhancements or marketing initiatives.

3.4. Sentiment Distribution by Name and Score

The most significant sentiments that were identified are

I. **Good Quality (Sentiment Score: 230):**

- Explanation: This sentiment indicates high praise for the quality of the product. Customers expressing "good quality" are likely satisfied with the durability, craftsmanship, or overall performance of the product. Positive sentiments regarding quality are crucial for building brand reputation and fostering customer loyalty.

II. **Good (Sentiment Score: 748):**

- Explanation: The sentiment "good" reflects overall satisfaction or approval of the product. Customers using this term are generally content with their purchase and perceive the product positively. "Good" sentiments contribute to positive brand perception and can drive repeat purchases and positive word-of-mouth referrals.

III. **Nice (Sentiment Score: 182):**

- Explanation: "Nice" signifies positive sentiment towards the product, albeit slightly less emphatic than "good" or "excellent." Customers using "nice" likely appreciate certain aspects of the product, such as its design, functionality, or aesthetics. While not as strong as other positive sentiments, "nice" still indicates a favorable impression.

IV. **Excellent (Sentiment Score: 25):**

- Explanation: "Excellent" conveys exceptionally positive feedback about the product. Customers using this sentiment are highly impressed with the product's performance, features, or value proposition. "Excellent" sentiments are highly desirable as they reflect outstanding customer satisfaction and endorsement of the product's superiority.

V. **Awesome (Sentiment Score: 28):**

- Explanation: "Awesome" indicates extremely positive sentiment and enthusiasm towards the product. Customers using this sentiment are likely to be highly satisfied and impressed with the product's attributes or capabilities. "Awesome" sentiments can have a strong impact on brand perception and customer advocacy.

VI. **Amazing (Sentiment Score: 22):**

- Explanation: "Amazing" reflects overwhelmingly positive sentiment and astonishment towards the product. Customers expressing "amazing" sentiments are exceptionally pleased with their experience and perceive the product as exceeding their expectations. "Amazing" sentiments are powerful endorsements that can attract new customers and enhance brand reputation.

VII. **Acceptable (Sentiment Score: 115):**

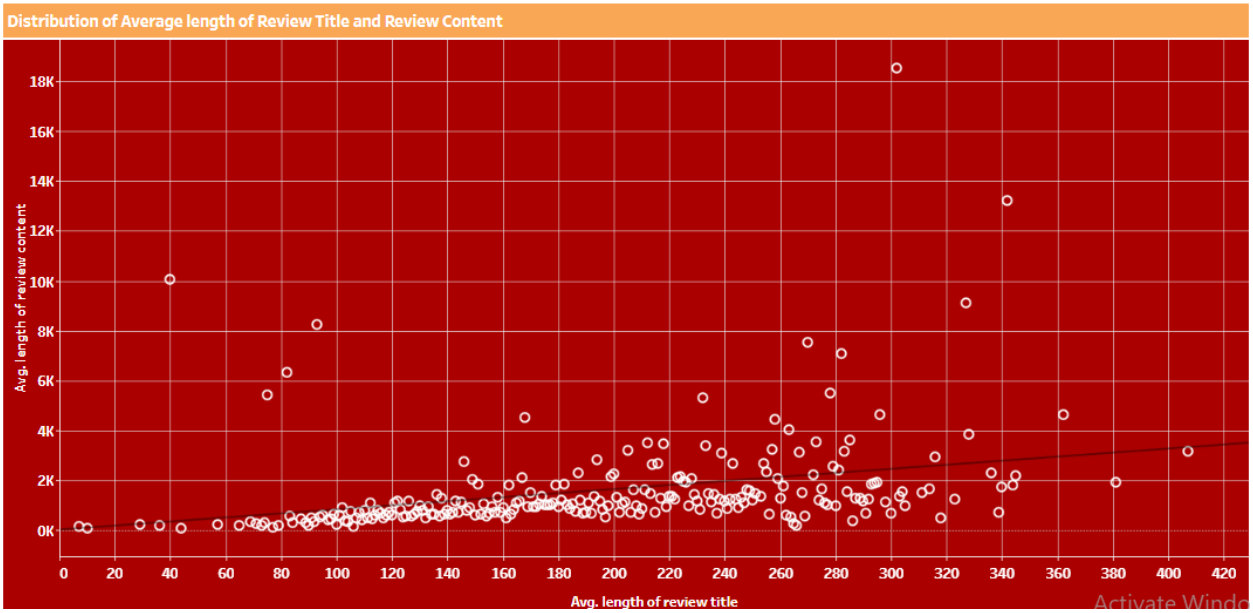
- Explanation: While not as effusive as some of the other sentiments, "acceptable" indicates a moderate level of satisfaction with the product. Customers using this sentiment find the product satisfactory or adequate, though they may not be as enthusiastic as those expressing stronger positive sentiments.

These sentiments provide valuable insights into customer perceptions, satisfaction levels, and areas of strength or improvement for the product. Understanding the nuances of customer feedback can help companies refine their products, enhance customer experiences, and drive business success.

4. Text Analysis:

4.1 Explore the length of review titles and review contents.

i. Average length of review content and review title



a. Trend Line Analysis:

The trend line starting at a low value and ending at a higher one suggests a gradual increase in the average length of review content across the dataset.

This upward trend indicates that, on average, review content tends to become longer as you move through the dataset.

b. Tightly Clustered Data:

The majority of the data being tightly clustered between 0 to 5k in average length of review content and 0 to 350 in average title length indicates a common pattern among reviews.

The clustering within this range suggests a consistent behavior among reviewers in terms of the length of both review content and titles.

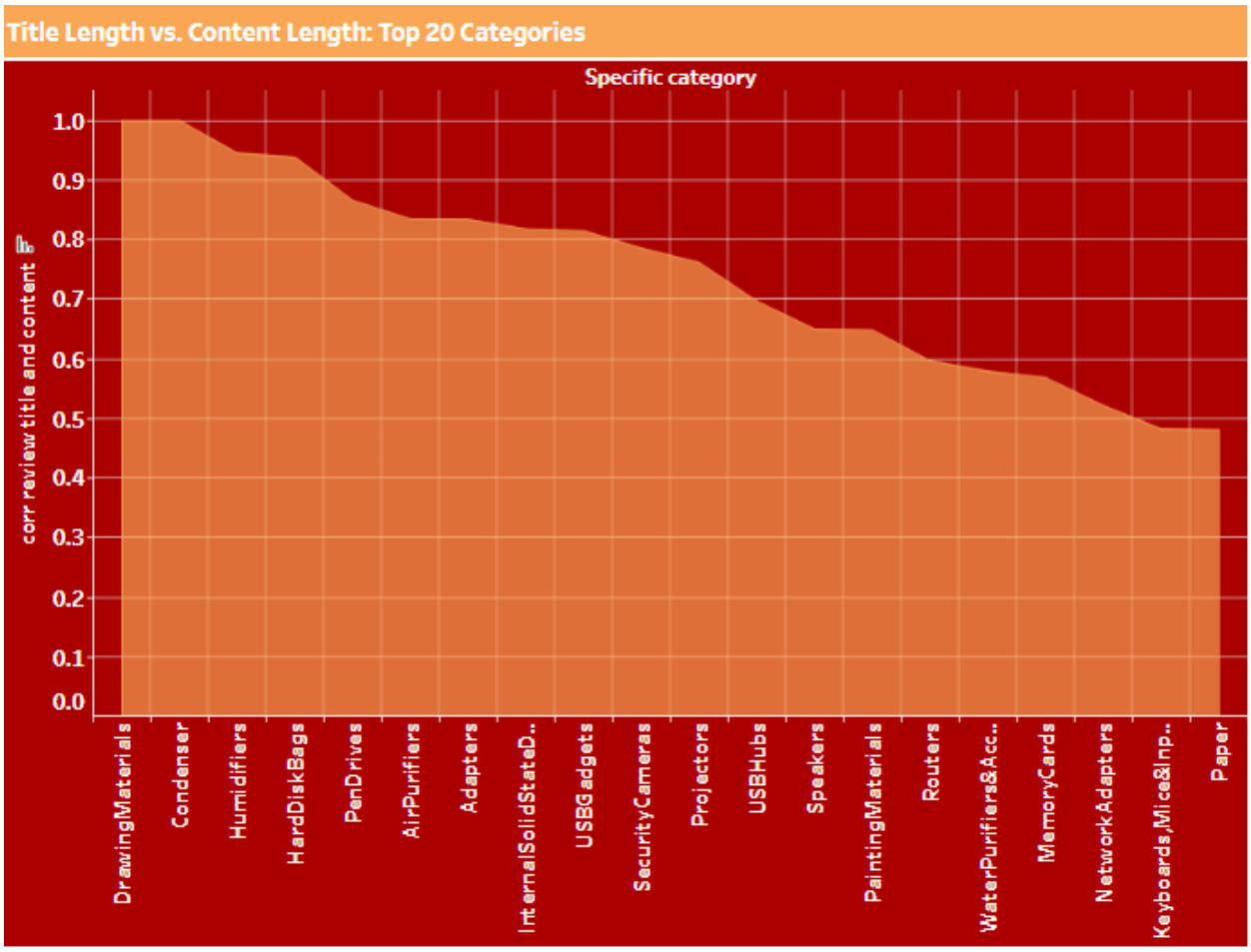
c. Possible Interpretations:

The clustering of data within the specified ranges could indicate a typical pattern in the lengths of review content and titles for the dataset under analysis.

It suggests that there may be certain conventions or expectations regarding the lengths of reviews and titles within this particular context.

The gradual increase in the average length of review content along the trend line may signify an evolving trend or a shift towards providing more detailed feedback over time.

ii. Correlation between review title length and review content length against specific category



a. Strong Positive Correlation:

Categories with correlation coefficients close to 1 indicate a strong positive correlation between review title and content lengths. This means that as the length of review titles increases, the length of review contents also tends to increase consistently.

Examples include Drawing Materials, Condenser, and Humidifiers. For these categories, longer review titles are likely to be accompanied by longer review contents.

b. Moderate Positive Correlation:
Categories with correlation coefficients between 0.5 and 0.8 suggest a moderate positive correlation between review title and content lengths. While not as strong as the previous category, there is still a noticeable tendency for longer review titles to be associated with longer review contents.
Examples include Pen Drives, Air Purifiers, and Speakers.

c. Weak Positive Correlation:
   Categories with correlation coefficients between 0 and 0.5 indicate a weak positive correlation between review title and content lengths. Although there is a positive relationship, it's not as consistent or pronounced compared to stronger correlations.
   Examples include Mobile Accessories, Accessories, and Small Kitchen Appliances.

d. Weak Negative Correlation:
   Categories with correlation coefficients between 0 and -0.5 suggest a weak negative correlation between review title and content lengths. This means that longer review titles may be associated with shorter review contents, and vice versa, but the relationship is not strong.
   Examples include Smartphones & Basic Mobiles, Audio & Video Accessories, and Rechargeable Batteries.

e. Strong Negative Correlation:
   Categories with correlation coefficients close to -1 indicate a strong negative correlation between review title and content lengths. In these categories, longer review titles are likely to be associated with shorter review contents, and vice versa.
   No categories in the provided list exhibit a strong negative correlation.

4.2 Most common words used in review titles and contents.

a. Overall Sentiment Distribution:

By categorizing sentiments based on their frequency, we can understand the overall sentiment distribution in the dataset. For example, sentiments like "good," "nice," "okay," and "acceptable" appear frequently, indicating that a significant portion of the feedback is positive or neutral.

b. Impact of Positive Sentiments:

Positive sentiments such as "good," "nice," "excellent," "awesome," and "amazing" are among the most frequently mentioned. This suggests that the majority of the feedback tends to be positive, indicating satisfaction or appreciation from customers.

c. Identification of Negative Sentiments:

Negative sentiments like "poor quality," "poor," "not worth," "not working," "disappointing," "bad quality," and "bad" also appear with noticeable frequency. This highlights areas of concern or dissatisfaction among customers, which should be addressed to improve overall customer experience and satisfaction.

d. Specific Issues Highlighted:

Certain negative sentiments, such as "fake," "broken," "expensive," and "high price," point towards specific issues or pain points experienced by customers. These insights can help identify areas for product improvement, pricing adjustments, or quality control measures.

e. Customer Feedback Trends:

Sentiments like "useful," "helpful," "recommended," and "reliable" indicate positive experiences and endorsements from customers. On the other hand, sentiments like "waste," "useless," "terrible," and "horrible" reflect negative experiences or dissatisfaction.

The frequency of these sentiments can provide insights into trends in customer feedback, allowing businesses to identify patterns and address recurring issues or capitalize on strengths.

f. Understanding Customer Perception:

Sentiments such as "fine," "tolerable," "adequate," and "decent" represent more neutral or lukewarm responses from customers. While these sentiments may not indicate extreme satisfaction or dissatisfaction, they still provide valuable feedback on areas where improvements or adjustments may be needed.

4.3 Sentiment analysis of reviews (positive, negative, neutral).

i. Sentiment Distribution:

The majority of sentiments expressed by customers fall into the "positive" category, with a total sentiment score of 1,342. This indicates that a significant portion of the feedback is positive, reflecting satisfaction or approval from customers.

While positive sentiments dominate, there are also instances of "negative" and "neutral" sentiments, with scores of 359 and 218, respectively. This suggests that there are areas of concern or dissatisfaction, as well as feedback that is neither explicitly positive nor negative.

ii. Intensity of Sentiments:

The sentiment scores provide insight into the intensity or magnitude of sentiments expressed by customers within each category. Higher sentiment scores in the "positive" category indicate stronger positive feelings or experiences shared by customers.

Conversely, the sentiment scores for "negative" and "neutral" sentiments suggest varying degrees of dissatisfaction or ambivalence among customers, with potentially different levels of impact on overall customer satisfaction and brand perception.

iii. Customer Engagement and Satisfaction:

The predominance of positive sentiments suggests that customers generally have positive experiences or perceptions of the product or service. This can be indicative of high levels of customer satisfaction, engagement, and loyalty.

However, the presence of negative and neutral sentiments highlights areas for improvement or attention. Analyzing the underlying reasons for negative and neutral feedback can help identify specific issues, pain points, or gaps in the customer experience that need to be addressed.

iv. Actionable Insights:

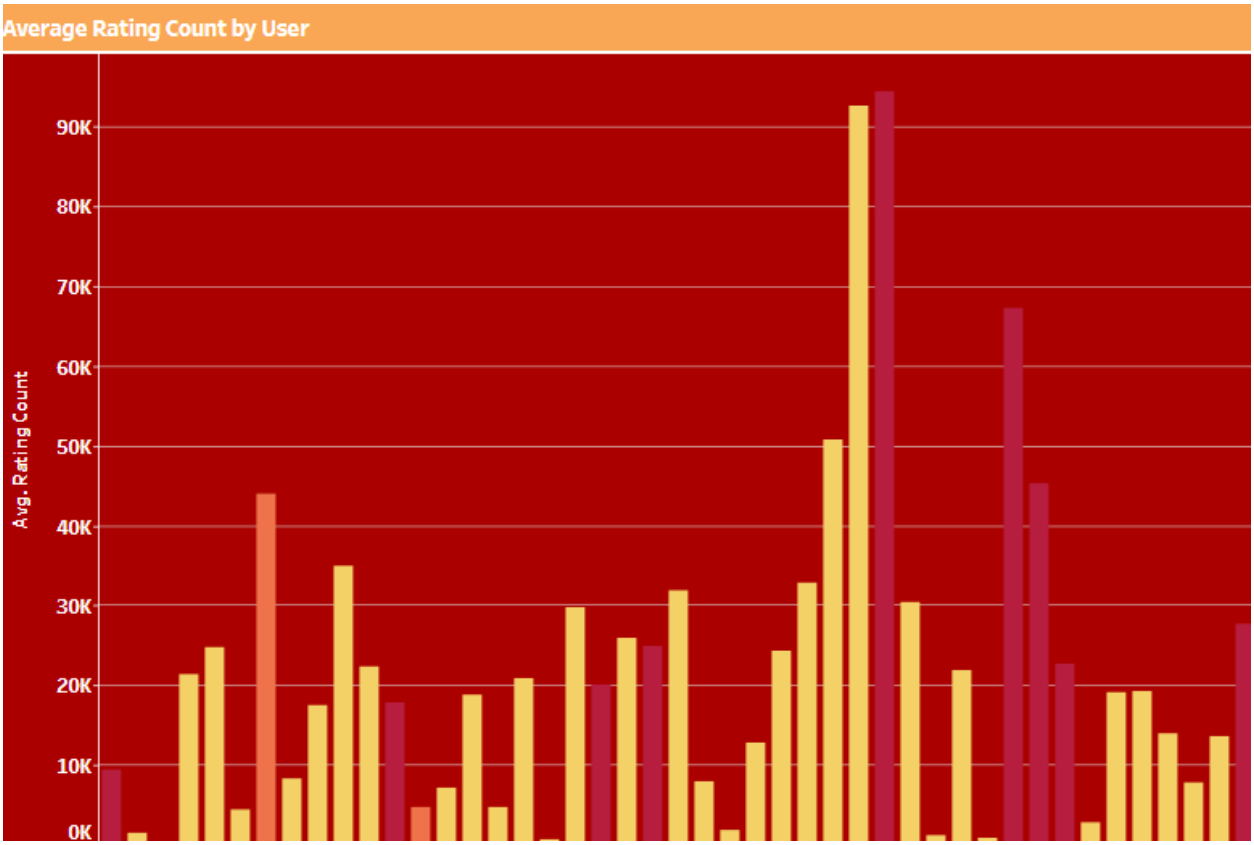
Businesses can use these insights to prioritize actions and initiatives aimed at amplifying positive feedback, addressing concerns raised in negative feedback, and enhancing overall customer satisfaction.

Understanding the distribution and intensity of sentiments can inform strategic decision-making processes related to product development, marketing strategies, customer service enhancements, and reputation management.

v. Trends and Patterns:

Monitoring changes in sentiment distribution and scores over time can reveal trends and patterns in customer feedback. Identifying shifts in sentiment can help businesses stay agile and responsive to evolving customer needs and preferences.

5. Feature engineering



5.1 User engagement

a. **High Engagement Users:**

- These users have a high average rating count, indicating active engagement with the platform.
- They are likely frequent users who provide feedback or ratings regularly.
- Targeted marketing or loyalty programs can be directed towards these users to maintain their engagement and loyalty.
- Analyzing their preferences and behavior can provide valuable insights into improving user experience and product offerings.

b. **Medium Engagement Users:**

- These users have a moderate average rating count, indicating some level of engagement with the platform.
- They may not be as active as high engagement users but still contribute to the platform's feedback.
- Strategies can be implemented to encourage these users to increase their engagement level, such as personalized recommendations or incentives for providing feedback.

- Understanding their usage patterns and preferences can help in tailoring marketing campaigns and improving user satisfaction.

c. **Low Engagement Users:**

- These users have a low average rating count, suggesting minimal engagement with the platform.
- They may be occasional users or passive observers who do not actively participate in providing feedback.
- Targeted efforts are required to re-engage these users and encourage them to provide feedback or interact more with the platform.
- Understanding the reasons behind their low engagement, such as usability issues or lack of incentives, can help in designing strategies to increase their involvement.

d. **Overall Insights:**

- High engagement users constitute a significant portion of the top 100 users, indicating a strong user base actively contributing to the platform.
- Medium engagement users also play a crucial role in providing feedback and contributing to the platform's success.
- Low engagement users represent an opportunity for growth and improvement, as efforts can be directed towards increasing their participation and feedback.
- Analyzing user engagement levels helps in understanding user behavior, identifying trends, and making data-driven decisions to enhance user satisfaction and platform performance.

5.2 User activity level

1. **High Engagement Users:**

    - Users in this category have a high average rating count, indicating active engagement with the platform.
    - Some notable high engagement users include Wraith, Vijayan C V, venkatesh kg, Siddharth patnaik, Satheesh Kadiam, Prasad Pavithran, Pavan A H, Neeraj Vishwakarma, Meghnad, Manoj maddheshiya, Manav, Mahantesh, livin sebi, Kapil kumar, Indro, and harpreet.
    - These users may be frequent buyers or reviewers who provide valuable feedback to other users.

2. **Medium Engagement Users:**
    - Users in this category have a moderate average rating count, suggesting some level of engagement with the platform.
    - Notable medium engagement users include Rahul Singh Rauthan, prateeq, Prashant, Omkar dhale, Jayesh, Hremant, AV, Binu, ASR, and Anonymous.
    - They may not be as active as high engagement users but still contribute to the platform's feedback and community.
3. **Low Engagement Users:**
    - Low engagement users have a low average rating count, indicating minimal engagement with the platform.
    - Users such as Ayush, ARDKN, and Amazon Customer fall into this category.
    - These users might be occasional buyers or observers who do not actively participate in providing feedback.
4. **Insights:**
    - The top users with high engagement levels contribute significantly to the platform's content and user experience.
    - Medium engagement users, although not as active as high engagement users, still play a vital role in providing feedback and contributing to the platform's success.
    - Low engagement users represent an opportunity for growth, and efforts can be directed towards increasing their participation and feedback.
    - Analyzing user engagement levels helps in understanding user behavior, identifying trends, and making data-driven decisions to enhance user satisfaction and platform performance.

6 Insights and Conclusions:

6.1 Key Insights:

i. User Engagement and Sentiment Analysis:

- High engagement users significantly contribute to the platform's content and user experience, indicating a strong user base actively involved in providing feedback and contributing to the platform's success.
- Positive sentiments dominate customer feedback, reflecting overall satisfaction, but the presence of negative and neutral sentiments underscores areas for improvement.
- Correlation analysis between review attributes reveals patterns in user behavior and preferences, with certain product categories exhibiting strong positive correlations between review title and content lengths.

ii. Pricing and Revenue Trends:

- The distribution of discounted prices varies widely across product categories, with certain categories experiencing occasional high discounts, potentially influencing customer purchase behavior.
- Products with higher ratings tend to have lower discount percentages, suggesting a correlation between perceived quality and pricing strategy.
- Categories like Smartphones & Basic Mobiles, Televisions, and Small Kitchen Appliances demonstrate high average revenue per unit, indicating potential high-profit segments for Amazon.

6.2 Recommendations:

i. Enhance Customer Engagement:

- Focus on engaging low engagement users by incentivizing participation and gathering feedback through targeted campaigns or rewards programs.
- Implement strategies to address areas of concern highlighted by negative and neutral sentiments, such as improving product quality, customer service, or addressing common pain points.

ii. Optimize Pricing and Discount Strategies:

- Analyze pricing and discount trends to identify opportunities for adjusting pricing strategies, especially for categories with low average revenue per unit.
- Ensure that discounts are aligned with customer preferences and perceived value to maintain customer satisfaction and brand reputation.

iii. Improve Product Offerings:

- Invest in product quality enhancements and innovation in categories with moderate average revenue per unit but low average rating counts to improve customer satisfaction and engagement.
- Leverage customer feedback and market insights to identify emerging trends and preferences, guiding product development and assortment planning.

iv. Strengthen Market Position:

- Capitalize on the popularity of high-revenue categories like Smartphones & Basic Mobiles and Small Kitchen Appliances by expanding product offerings, enhancing customer service, and strengthening brand presence.
- Explore partnerships or collaborations with trusted brands to offer exclusive products or promotions, attracting new customers and increasing market share.

Overall, by addressing the identified opportunities and challenges, Amazon can further solidify its position as a leading online marketplace, driving customer satisfaction, loyalty, and sustainable growth.

7. Conclusion

7.1 Recap of the EDA Process and Findings:

- The exploratory data analysis (EDA) process delved into various aspects of the dataset, including user engagement, sentiment analysis, correlation between review elements, pricing trends, and revenue performance across product categories.
- Key findings revealed patterns such as strong positive correlations between review title and content lengths in certain categories, dominance of positive sentiments with areas for improvement identified through negative and neutral feedback, and variability in pricing strategies and revenue performance across different product segments.

7.2 Significance of the Analysis and Potential Next Steps:

- The analysis provided valuable insights into user behavior, customer sentiment, and market dynamics, offering actionable recommendations to enhance user engagement, address feedback, and optimize pricing strategies.

- Moving forward, potential next steps include implementing targeted initiatives to increase user engagement among low-engagement segments, addressing specific pain points identified through sentiment analysis, and refining pricing strategies to maximize revenue potential.
- Additionally, ongoing monitoring and analysis of user interactions, sentiment trends, and market dynamics will be crucial for maintaining a competitive edge and adapting strategies to evolving customer preferences and market conditions.

In conclusion, the EDA process has provided valuable insights that can inform strategic decision-making and drive improvements in user satisfaction, platform performance, and revenue growth. Continued analysis and proactive measures will be essential for sustaining success and achieving long-term objectives in the ever-changing landscape of the digital marketplace.