Course 5: Data Mining Summative Assignment

Market Basket Analysis of Amazon by Customer Segmentation and Shopping Patterns

1. Student Name: Prakhar Sharma
2. Candidate Registration Number: 1000260
3. CRS Name: Artificial Intelligence
4. Course Name: IDAI1
5. School Name: Ryan Global School, Kharghar

Project Scope: Amazon E-commerce Data Analysis

Objective:

The primary objective of this project is to **analyze Amazon's e-commerce** data to:

1. **Uncover Customer Shopping Behaviors** – Understand customer purchasing patterns, preferences, feedback and engagement.
2. **Segment Customers** – Group customers based on buying patterns for targeted marketing strategies.
3. **Identify Product Relationships** – Discover associations between products for effective cross-selling and bundling.Identification of products frequently bought together, supporting strategic product placements and promotions.
4. Improve marketing strategies, product recommendations, and customer satisfaction.
5. **Interactive Dashboard:** Deployment of an interactive Streamlit dashboard for real-time data exploration and decision-making.

Purpose of Each Analysis Step:

1. **Exploratory Data Analysis (EDA):**
    - **Purpose:**
        - To understand the dataset's structure, distribution, and key features.
        - To identify trends, patterns, and potential outliers.
    - **Contribution:**
        - Informs data cleaning, feature selection, and model development.

        - Provides foundational insights into pricing strategies, product demand, and customer ratings.
2. **Customer Segmentation:**
    - **Purpose:**
        - To group customers based on their shopping behaviors such as spending patterns, product preferences, and feedback for targeted marketing.
    - **Contribution:**
        - Helps in targeted marketing, personalized recommendations, and customer retention strategies.
        - Enables strategic decision-making for promotional campaigns and product recommendations.
3. **User Behavior Analysis:**
    - **Purpose:**
        - To analyze customer reviews, ratings, and feedback to understand user sentiments and engagement(for better offerings)
    - **Contribution:**
        - Identifies customer satisfaction drivers and pain points.
        - Supports product improvement, reputation management, and personalized customer interactions.
4. **Association Rule Mining:**
    - **Purpose:**
        - To identify frequent itemsets and product combinations that are commonly bought together ie. Identifies product relationships for bundling
    - **Contribution:**
        - Facilitates effective cross-selling, product bundling, and inventory management.
        - Enhances customer experience through personalized product recommendations.
5. **Deployment with Streamlit:**
    - **Purpose:**
        - To present the findings interactively using a user-friendly dashboard.
    - **Contribution:**
        - Enables real-time exploration of data insights for stakeholders.
        - Facilitates informed, data-driven decisions to enhance marketing strategies and product offerings.

Strategic Impact:

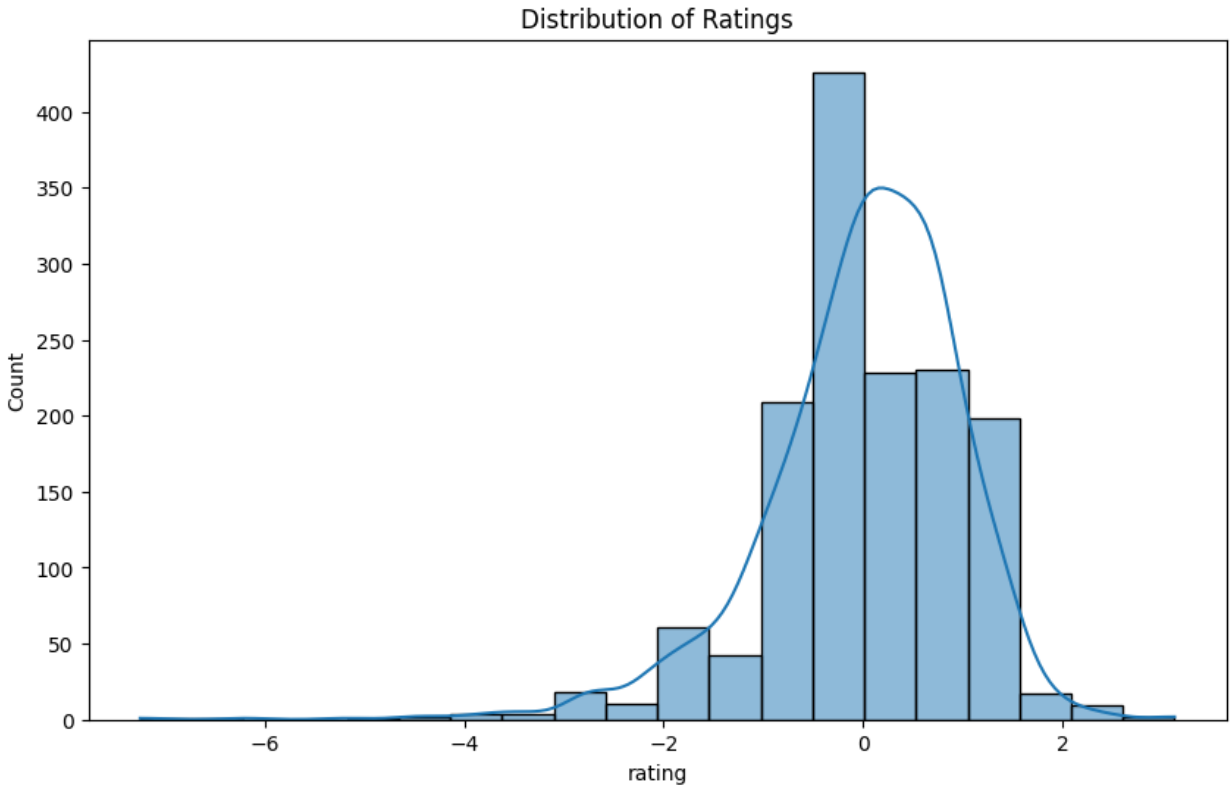
1. **Marketing Optimization:**

    - Develop targeted marketing strategies based on customer segments.
    - Enhance promotional effectiveness by identifying high-demand products and popular categories.
2. **Product Recommendations:**
    - Utilize product associations to personalize recommendations.
    - Increase sales through effective cross-selling and bundling strategies.
3. **Customer Satisfaction:**
    - Improve customer satisfaction by understanding preferences and addressing feedback.
    - Enhance user engagement through personalized experiences and recommendations.
4. **Business Decision Support:**
    - Provide actionable insights to support strategic decision-making.
    - Facilitate dynamic exploration and reporting through the interactive Streamlit dashboard.

Summary:

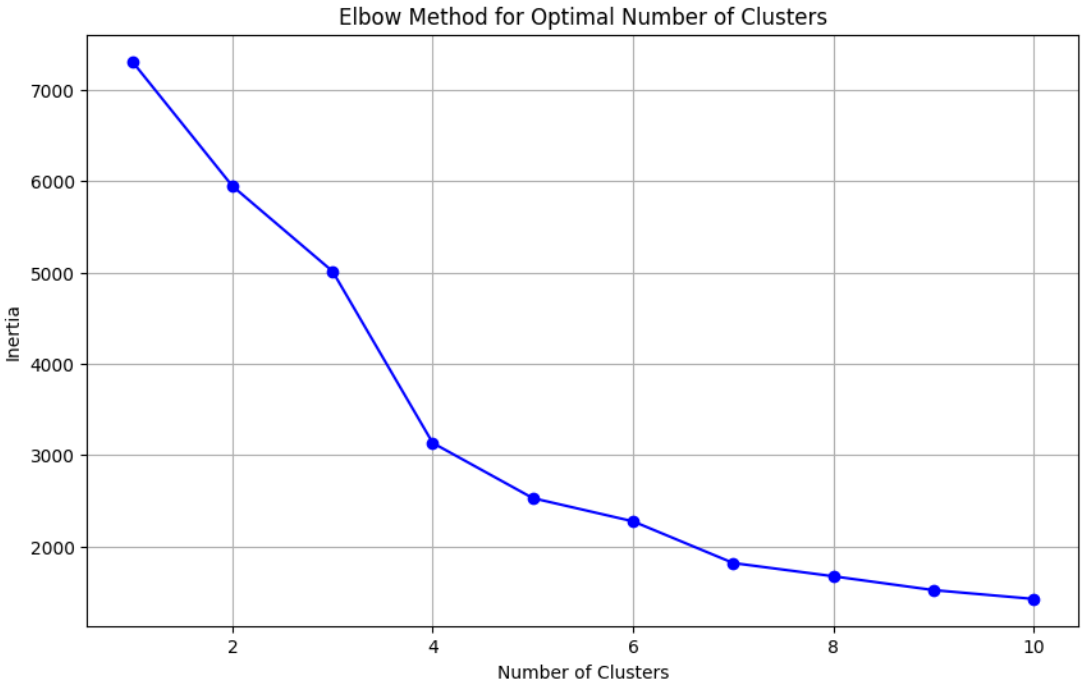
This project aims to transform Amazon's e-commerce data into actionable insights that drive personalized marketing strategies, optimized product recommendations, and enhanced customer satisfaction. By systematically analyzing customer behaviors, segmenting customers, and identifying product relationships, this project empowers data-driven decision-making and strategic growth.

Data Cleaning and Preprocessing:



Key Observations:

1. **Distribution Shape and Skewness:**
   - The distribution is slightly **left-skewed (negatively skewed)**, with a longer tail on the negative side.
   - Most of the ratings are clustered between **-1 and 1**, indicating a predominance of neutral or slightly positive sentiments.
2. **Central Tendency and Peak:**
   - The peak is around **0**, signifying that a majority of ratings are neutral.
   - This suggests that customers neither strongly liked nor disliked the products, indicating average satisfaction.
3. **Outliers and Spread:**
   - There are some extreme negative ratings (below -4), which may represent a subset of highly dissatisfied customers.
   - The positive end has a sharper decline, showing fewer highly positive ratings compared to the neutral and slightly negative ones.
4. **Normality and KDE Curve:**
   - The KDE (Kernel Density Estimate) curve approximates a normal distribution but is skewed left.
   - This indicates that while most ratings are around the mean (0), there is a heavier tail on the negative side.



Elbow Method Analysis for Optimal Clustering

1. Understanding the Graph

- **X-axis (Number of Clusters):** Represents the different numbers of clusters considered in the K-Means clustering algorithm.
- **Y-axis (Inertia):** Measures the sum of squared distances between data points and their assigned cluster centroids (also known as the within-cluster sum of squares).
- **Curve Pattern:**
    - As the number of clusters increases, inertia decreases since clusters become smaller and better fit the data.
    - However, after a certain number of clusters, the rate of inertia reduction **diminishes**, forming an **elbow-like shape**.

2. Identifying the Optimal Number of Clusters

- The "elbow" point is where adding more clusters **does not significantly reduce inertia** anymore, meaning additional clusters provide **diminishing returns** in improving clustering quality.
- From the graph, the elbow appears to be around **4 clusters** (as inertia drops sharply up to this point, then levels off).
- This suggests that **4 clusters** might be the optimal choice for segmenting the data.

3. Business Implications & Use Cases

- **Customer Segmentation:** If this data is from an e-commerce dataset, the clusters may represent different types of customers (e.g., budget shoppers, premium buyers, frequent shoppers, one-time buyers).
- **Product Categorization:** If clustering is applied to products, the groups may represent different pricing categories, popularity levels, or demand trends.
- **Fraud Detection:** If used in fraud analysis, clusters may differentiate normal user behavior from potential fraudulent activity.

Observations:

1. **Discounted Price and Ratings:**
    - Most data points are concentrated at low discounted prices (close to 0) and ratings near 0.
    - A sparse distribution is observed as the discounted price increases.
    - Negative ratings are more frequently seen for lower discounted prices.
2. **Customer Segments:**

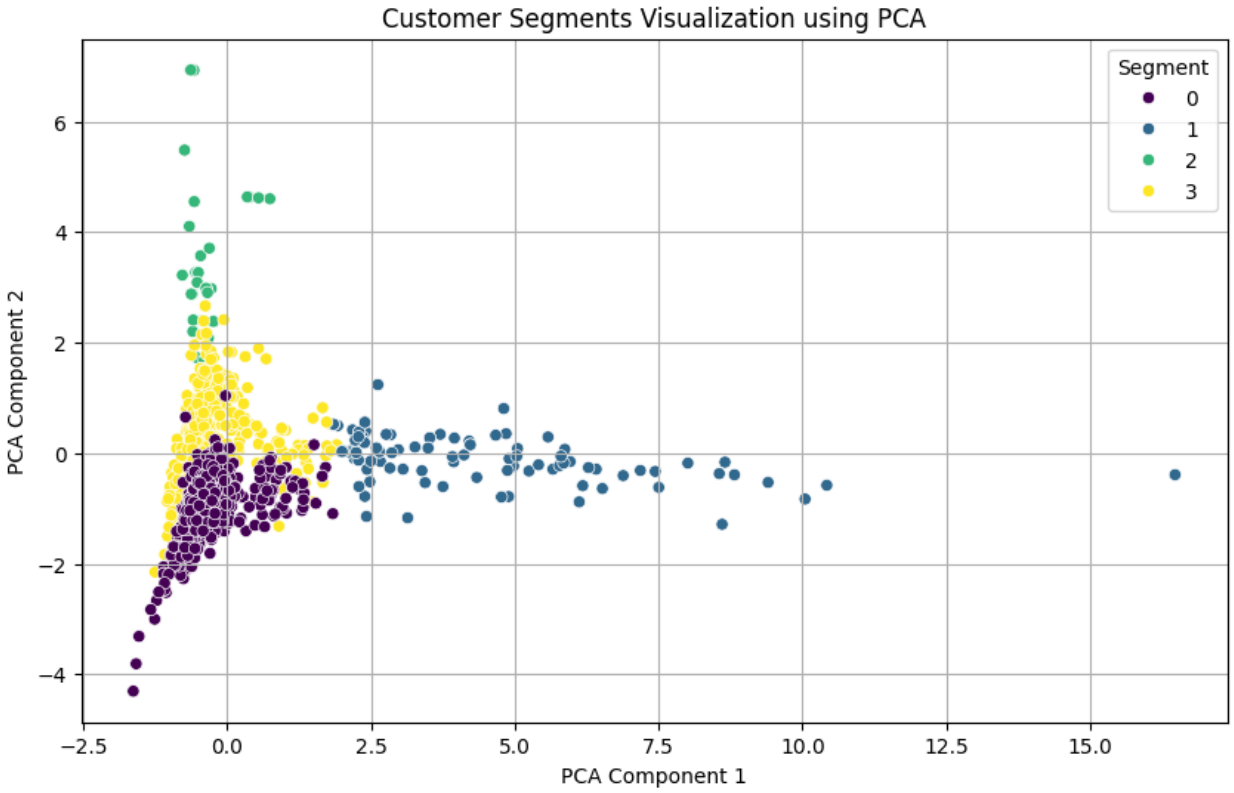
    - Segments `0` and `1` dominate at lower discounted prices and ratings.
    - Segments `2` and `3` appear more prominent for higher discounted prices.
3. **Rating Count:**
    - The majority of data points have small-sized dots, indicating low rating counts.
    - Only a few points have a significant number of ratings (larger dots), mostly for discounted prices near 0.

Analysis:

1. **Customer Behavior:**
    - Segment `0` customers tend to interact more with low-discounted items, but their ratings are clustered near 0 or negative.
    - Higher-segment customers (e.g., `3`) seem more interested in higher discounted prices and give ratings closer to positive values.
2. **Pricing and Feedback:**
    - Higher discounted prices seem to receive better feedback, albeit with fewer interactions.
    - Lower discounted prices receive a mixed range of ratings but dominate the dataset, suggesting they drive more customer interactions.
3. **Segments Differentiation:**
    - Segment-based differentiation can provide valuable insights for targeted marketing or promotions.

Recommendations:

1. **Tailored Promotions:**
    - For segment `0`, focus on low-priced items but address why their ratings are neutral/negative.
    - For segment `3`, emphasize higher-quality or higher-discount items since they show a preference for these.
2. **Rating Analysis:**
    - Investigate why certain segments (e.g., `0`) tend to give lower ratings and address potential dissatisfaction.
3. **Enhanced Engagement:**
    - Consider strategies to increase rating counts for higher discounted items



Analysis of Customer Segments using PCA

1. Overview of the Visualization

- The scatter plot represents **customer segmentation** using **Principal Component Analysis (PCA)**.
- PCA reduces the dimensionality of the dataset, allowing for better visualization of customer clusters.
- Each point corresponds to a customer, positioned according to its transformed features.

2. Key Observations

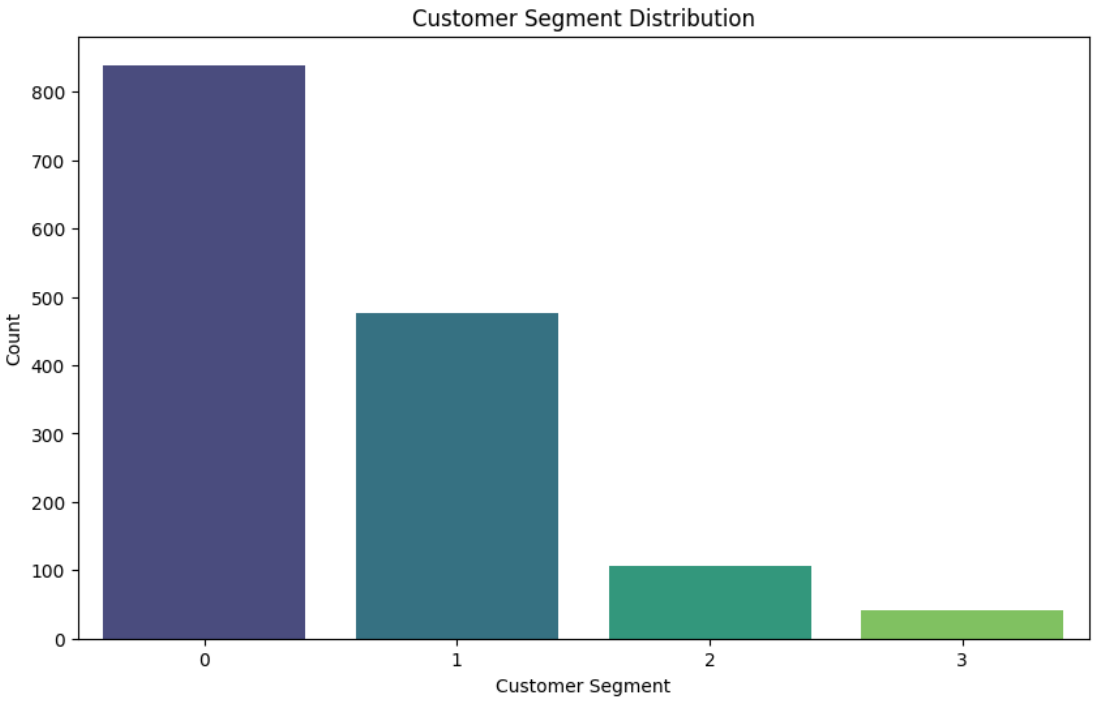
- **Clusters (Segments 0, 1, 2, 3)**
  - **Purple (Segment 0)**: Densely packed, mostly near the origin, indicating a core group of customers with similar purchasing behavior.
  - **Blue (Segment 1)**: More dispersed, spreading horizontally along PCA Component 1, possibly representing diverse spending patterns.
  - **Green (Segment 2)**: Positioned towards the upper-left, indicating a unique subset of customers with distinct behavior.
  - **Yellow (Segment 3)**: Overlapping with purple but slightly spread out, suggesting customers with some behavioral similarities but minor variations.

3. Insights from Clustering

- **Purple and Yellow clusters (Segments 0 & 3)** seem closely related, possibly indicating similar purchasing trends with minor distinctions.
- **Blue cluster (Segment 1)** extends significantly along the PCA Component 1 axis, suggesting a wider range of customer activity levels.
- **Green cluster (Segment 2)** is sparsely populated and slightly separated, indicating customers with unique characteristics (e.g., high-value customers or outliers).

4. Business Implications

- **Segment 0 & 3 (Core Customers)**: Targeted retention strategies, loyalty programs, and personalized promotions could enhance engagement.
- **Segment 1 (Diverse Customers)**: A flexible marketing approach may be needed, as they exhibit varied behavior.
- **Segment 2 (Outliers or High-Value Customers)**: Premium services, personalized recommendations, and VIP treatment could boost revenue.



Analysis of Customer Segment Distribution

1. Overview of the Visualization

- This bar chart represents the **distribution of customers across different segments**.
- The **x-axis** represents the customer segments (0, 1, 2, 3).
- The **y-axis** represents the count of customers in each segment.

2. Key Observations

- **Segment 3 (Green) has the highest number of customers (~850+).**
  - This suggests that the majority of customers fall into this segment.
- **Segment 0 (Dark Blue) is the second-largest group (~470 customers).**
  - A significant portion of the customer base belongs here.
- **Segment 1 (Teal) and Segment 2 (Light Green) are the smallest groups.**
  - Segment 1 has around **100 customers**.
  - Segment 2 has even fewer, likely **under 50 customers**.

Analysis:

1. **Segment `0`:**
   - This is the largest customer segment with over 800 customers.
   - Indicates that most customers fall into this group.
2. **Segment `1`:**
   - The second-largest group with approximately 500 customers.
   - A significant portion of the customer base also belongs to this segment.
3. **Segment `2`:**
   - A smaller group with fewer than 200 customers.
   - Represents a specialized or niche segment.
4. **Segment `3`:**
   - The smallest group with fewer than 100 customers.
   - Likely represents a highly specific or premium customer base.

Observations:

- The majority of customers are concentrated in **segments `0` and `1`**, indicating these segments are more dominant or broadly defined.
- **Segments `2` and `3`** are much smaller, possibly representing specialized, high-value, or less common customer categories.
- Segment `0` might require tailored strategies to manage its large size, while segments `2` and `3` may need focused attention to grow.

3. Business Implications

Segment 0:

- Dominates the customer base, requiring significant resources to serve and retain.
- Business should focus on optimizing operations to efficiently cater to this group.
- Risk: Overdependence on this segment could make the business vulnerable if its preferences or behavior change.
- Marketing and Sales Strategies- Focus on cost-effective, large-scale marketing campaigns to maintain dominance.Leverage loyalty programs and bulk offers to keep engagement high.
-

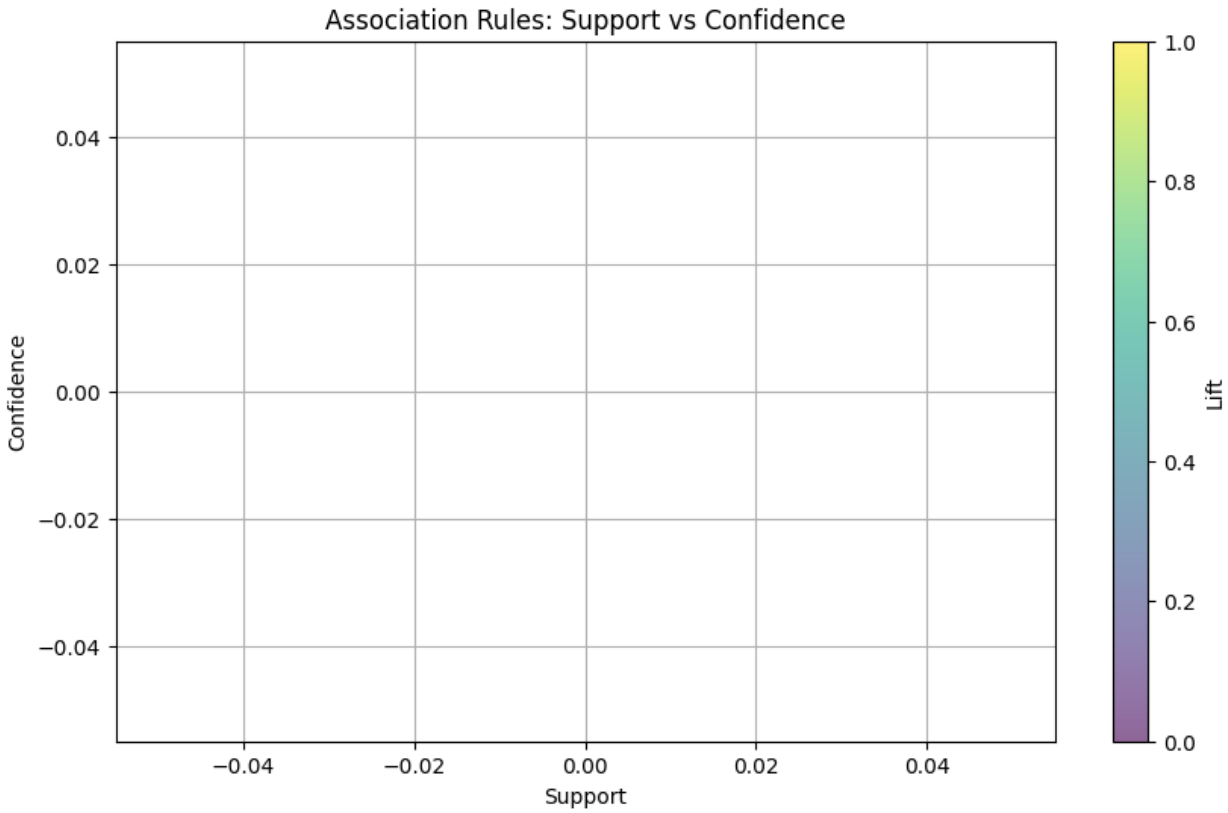
Segment 1:

- A strong secondary group; can be targeted for upselling and cross-selling.
- Investing in personalized marketing strategies could strengthen loyalty and revenue from this group.
- Marketing and Sales Strategies-Introduce mid-tier or customizable products to encourage higher spending.Use targeted digital advertising to attract more customers similar to this segment.

Segment 2 and 3:

- Represent smaller but potentially high-value or premium customer groups.
- Tailored experiences, exclusive offers, or specialized services for these segments could improve customer retention and attract more customers with similar preferences.
- These groups could act as a testing ground for new premium products or services.
- Marketing and Sales Strategies-Create exclusivity-driven campaigns (e.g., VIP memberships, premium services).Emphasize quality and value to attract and retain niche customers.

Association Rule Mining:



Analysis of Association Rules: Support vs Confidence Plot

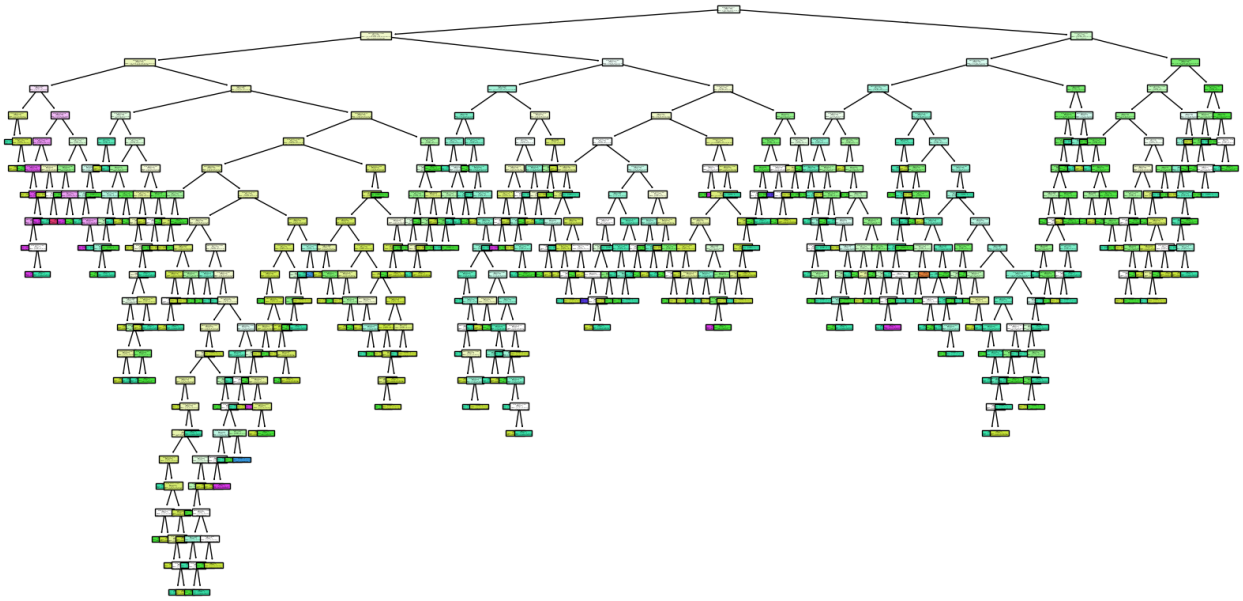
1. Overview of the Visualization

- This scatter plot is meant to visualize association rules by plotting **Support (x-axis) vs Confidence (y-axis)**.
- The **color bar represents Lift**, which indicates the strength of the association.
- However, the graph appears **empty**, meaning that no association rules were generated.

2. Key Observations

- The absence of data points suggests one of the following:
    - **No frequent itemsets met the minimum support threshold** → The dataset may have sparse transactions.
    - **No rules met the minimum confidence threshold** → The confidence level may be too high, filtering out all rules.

- **Data preprocessing issues** → Issues in encoding transactions or applying the Apriori/FP-Growth algorithm.
- **Incorrect parameter tuning** → The chosen support and confidence thresholds might be too restrictive.



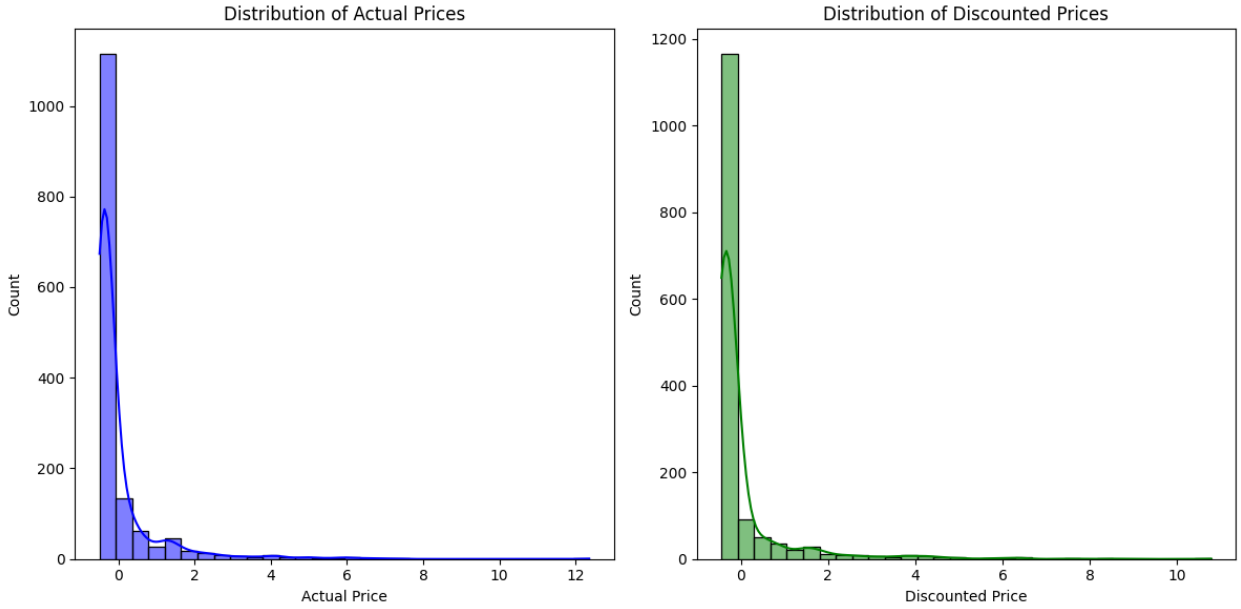
Key Observations:

1. **Tree Structure and Complexity:**
   - The tree is quite deep and complex, indicating the model is fitting a lot of rules to the training data.
   - This complexity may suggest **overfitting**, where the model is capturing noise rather than generalizable patterns.
2. **Feature Importance and Splits:**
   - The top nodes are the most important split criteria. If **actual price** or **discounted price** are the top splits, they are the most significant predictors.
   - The tree continues to split into smaller subsets, likely refining predictions for various price ranges or discount categories.
3. **Class Distribution and Prediction:**
   - The leaf nodes represent the final predictions or classes. The color and density of nodes show the distribution of predicted classes or values.

- Balanced colors across branches suggest a well-distributed model, while dominant colors might indicate class imbalance or bias.

4. **Model Accuracy and Performance:**
   - The depth and branching suggest high training accuracy but might result in poor generalization on new data due to potential overfitting.
   - If validation accuracy is much lower than training accuracy, **pruning** or **regularization** is recommended.

Exploratory Data Analysis (EDA):

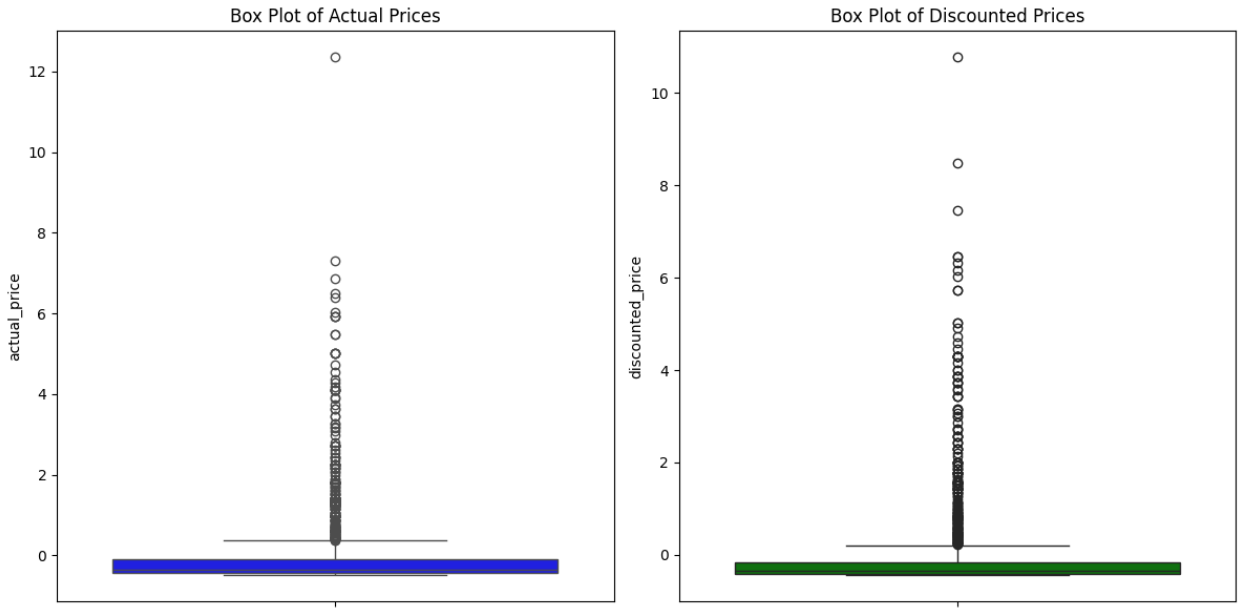


Analysis of Price Distributions:

1. **Skewness and Distribution Shape:**
   - Both **Actual Price** and **Discounted Price** distributions are **heavily right-skewed**, indicating that most products are priced at the lower end, with a few high-priced outliers.
   - This is typical in markets where affordable items are more common, while premium products are rare but significantly more expensive.
2. **Price Range and Outliers:**
   - **Actual Prices** go up to approximately **140,000**, but the majority are concentrated below **20,000**.
   - **Discounted Prices** reach around **80,000**, but most are below **10,000**, suggesting significant discounts on high-priced items.

- The long tails in both distributions indicate the presence of outliers, which can heavily influence model performance if not handled properly.

3. **Discount Analysis and Pricing Strategy:**
   - The leftward shift of the **Discounted Price** distribution compared to the **Actual Price** highlights effective discounting strategies to attract more customers.
   - The sharper drop-off in the discounted price curve suggests that higher-priced items are more aggressively discounted, possibly to clear premium inventory.
4. **Implications for Predictive Modeling:**
   - The skewness and outliers could **negatively impact model accuracy** if not properly addressed. Consider applying:
     - **Log Transformation:** To reduce skewness and stabilize variance.
     - **Outlier Treatment:** Capping extreme values or using robust models (e.g., XGBoost).
     - **Normalization or Standardization:** To improve model convergence and accuracy.



Box Plot Analysis of Actual and Discounted Prices:

1. **Outliers and Skewness:**
   - Both **Actual Price** and **Discounted Price** box plots show a **significant number of outliers** above the upper whisker, confirming the heavy right-skew observed in the histograms.

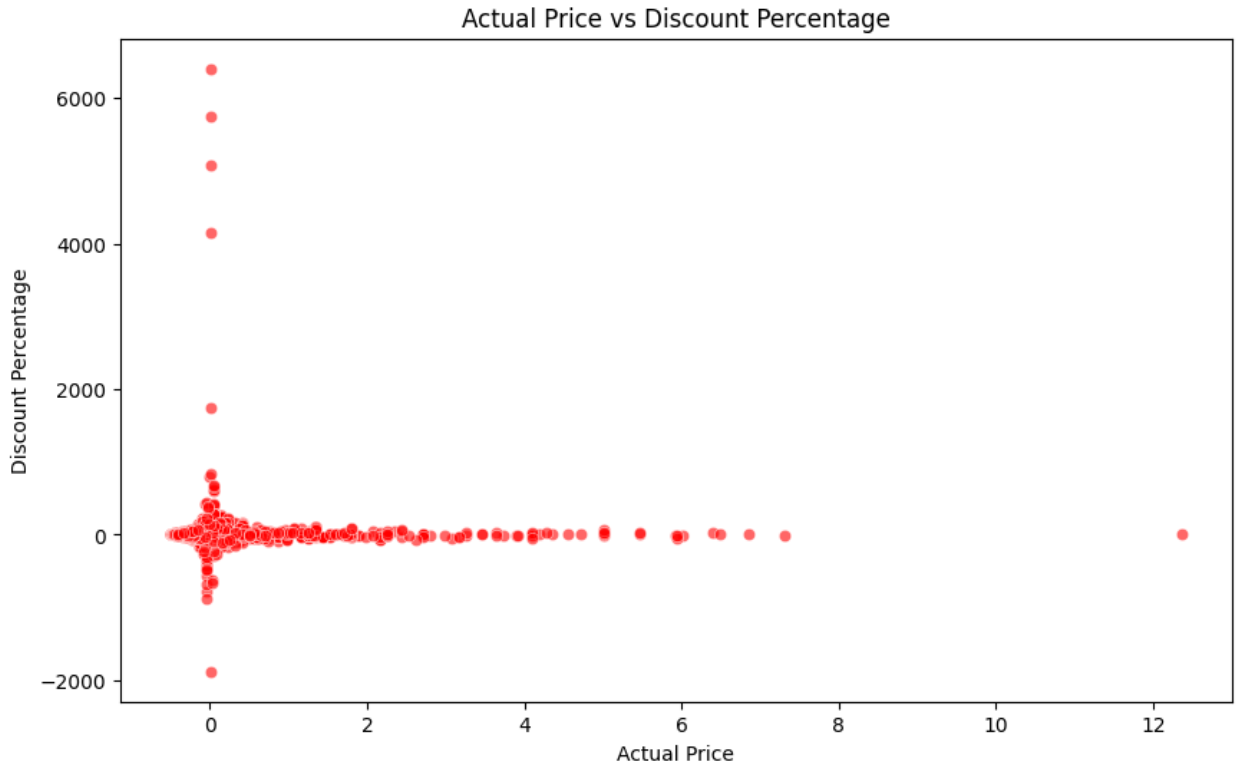
    - These outliers represent high-priced items, indicating a small portion of luxury or premium products.
2. **Price Range and Central Tendency:**
    - The interquartile range (IQR) for both distributions is **heavily concentrated near the lower end**, showing that most products are priced within a narrow range.
    - **Median Prices:**
        - Actual Prices: The median is relatively low compared to the overall range, indicating affordable products dominate the market.
        - Discounted Prices: The median is even lower, highlighting the impact of discounts on making products more accessible.
3. **Discount Impact:**
    - The compression of the IQR in the **Discounted Prices** compared to the **Actual Prices** suggests a **reduction in price variance** due to discounts, making prices more uniform.
    - This aligns with a marketing strategy aimed at increasing sales volume by appealing to price-sensitive customers.
4. **Extreme Outliers:**
    - In **Actual Prices**, extreme outliers exceed **140,000**, whereas in **Discounted Prices**, the maximum is around **80,000**, indicating substantial markdowns on premium products.
    - These extreme values may distort mean calculations and influence predictive model performance.

Scatter Plot Analysis: Actual Price vs. Discounted Price

1. **Positive Correlation:**
    - There is a **strong positive correlation** between **Actual Price** and **Discounted Price**, indicating that higher-priced products tend to have higher discounted prices as well.
    - This suggests a consistent discounting strategy proportional to the actual price, possibly maintaining perceived value across all price tiers.
2. **Linear Trend with Variability:**
    - The points generally follow a **linear trend**, but the spread increases with higher prices, showing **greater variance in discounting strategies** for expensive products.
    - This could indicate customized pricing or targeted promotions for high-end products.
3. **Cluster Analysis:**
    - A **dense cluster** is visible in the lower price range (below 20,000), showing that most products are affordably priced and have relatively small discounts.
    - The data is more **dispersed at higher price ranges**, reflecting diverse discounting strategies or product categories.
4. **Outliers and Anomalies:**

- A **notable outlier** is present at the upper end (Actual Price ≈ 140,000 and Discounted Price ≈ 80,000), suggesting a luxury item with a substantial discount.
- A few points show **lower discounts** than expected for their price range, potentially indicating **premium pricing or minimal promotions**.

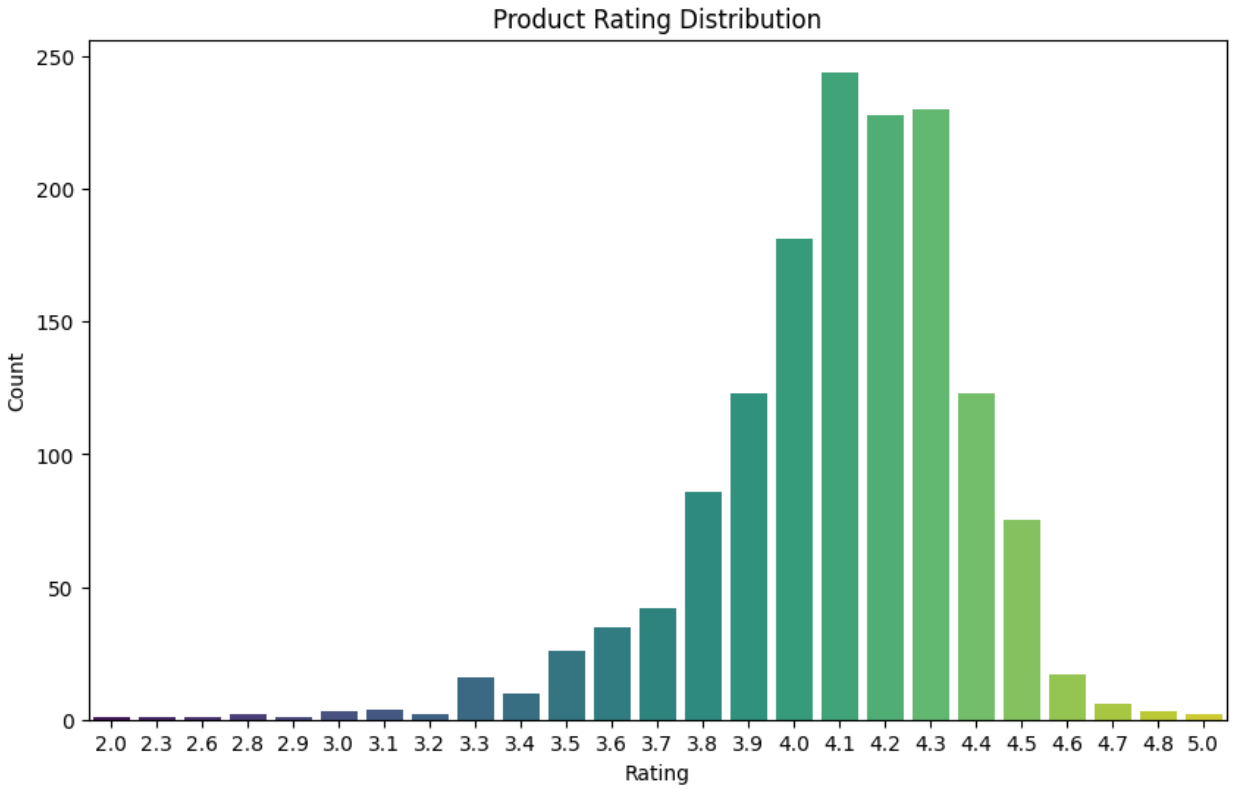
5. **Implications for Pricing Strategy:**
   - The correlation supports **predictive modeling** for discount optimization, potentially using **linear regression** or **gradient boosting** techniques.
   - The variability in discounts suggests opportunities for **dynamic pricing strategies** or **personalized promotions** based on customer segmentation.
6. **Next Steps for Analysis:**
   - Perform **correlation analysis** (e.g., Pearson or Spearman coefficient) to quantify the relationship.
   - Use **log transformation** to stabilize variance for predictive modeling.
   - Segment the data into price tiers to investigate **differential discount strategies**.



Scatter Plot Analysis: Actual Price vs. Discount Percentage

1. **Inverse Relationship with Price:**
   - Higher-priced items generally have **lower discount percentages**, indicating a strategy to preserve premium brand value.

    - Lower-priced products show a **wide range of discount percentages**, suggesting aggressive discounting to attract price-sensitive customers.
2. **Dense Cluster at Low Prices:**
    - A **dense cluster** is observed at lower actual prices (below 20,000) with a **high variation in discount percentages** (0% to 80%).
    - This indicates **price-based segmentation**, where discounts are used strategically to boost volume sales in lower price brackets.
3. **Sparse Distribution for High Prices:**
    - As the actual price increases, the points become more **sparse and scattered**, reflecting a **conservative discounting strategy** for expensive items.
    - This could be due to **profit margin protection** or maintaining a **premium brand perception**.
4. **Outliers and Anomalies:**
    - A few outliers show **high discount percentages** for high-priced products, possibly indicating **inventory clearance** or **special promotions**.
    - One extreme outlier (Actual Price ≈ 140,000 with a high discount percentage) suggests a **deep discount on a luxury item**, likely for promotional purposes.
5. **Implications for Pricing and Marketing:**
    - The pattern suggests a **tiered discounting strategy**, where discounts are used aggressively for lower-priced products but conservatively for high-end items.
    - This aligns with strategies to **maximize revenue while preserving brand equity**.
6. **Next Steps for Analysis:**
    - Perform a **correlation analysis** to quantify the relationship between price and discount percentage.
    - Segment the data by **price tiers** to explore targeted discount strategies.
    - Conduct a **regression analysis** to predict discount percentages based on product pricing and other features (e.g., category, seasonality).

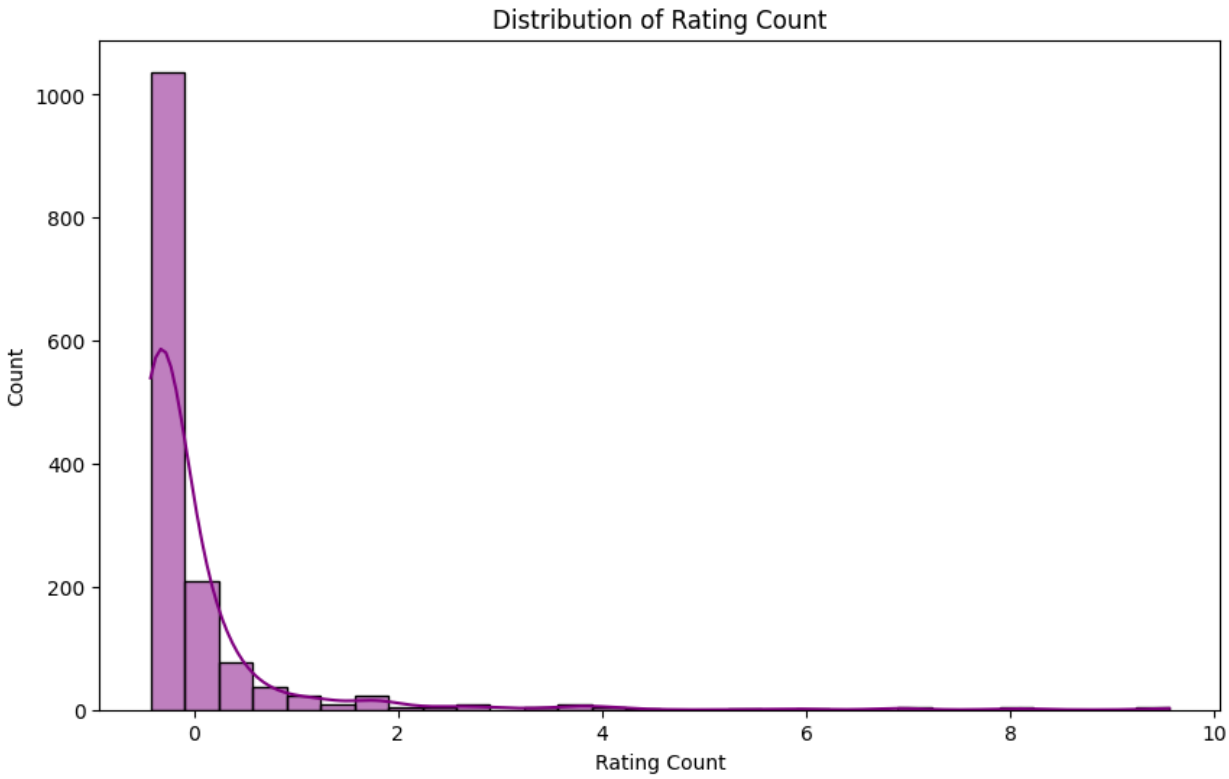


Histogram Analysis: Product Rating Distribution

1. **Positive Skewness and High Average Rating:**
    - The distribution is **positively skewed** with the majority of ratings concentrated between **4.0 and 4.3**, indicating a general trend of positive customer feedback.
    - This suggests **high customer satisfaction** and a **positive brand perception** for most products.
2. **Peak at 4.0 - 4.2:**
    - The **highest frequency** of ratings is observed between **4.0 and 4.2**, reflecting a tendency for customers to rate products favorably but not at the maximum level.
    - This could be due to **high expectations** or **minor issues** preventing perfect ratings.
3. **Low Frequency of Extreme Ratings:**
    - Very few products received ratings below **3.0** or above **4.7**, indicating **consistent product quality** with minimal dissatisfaction or exceptional delight.
    - The **scarcity of 5.0 ratings** suggests customers are critical even when satisfied, possibly due to a desire for improvement or cautious optimism.
4. **Long Tail towards Lower Ratings:**

- The **long tail** on the lower end (below 3.5) suggests a **small subset of underperforming products**, which may require quality improvements or targeted marketing strategies.

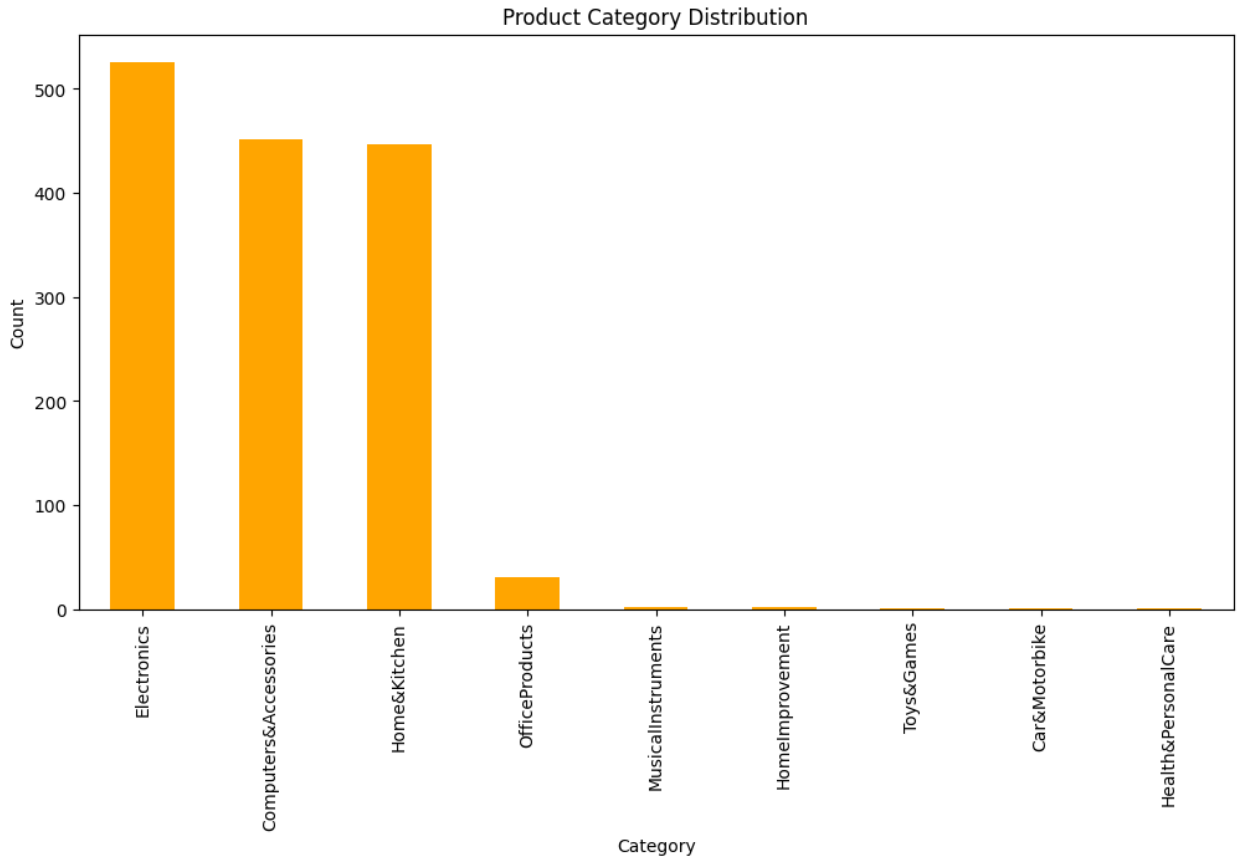
5. **Implications for Product Strategy:**
   - Products rated **below 3.5** should be **investigated for quality issues** or **negative feedback** to identify improvement opportunities.
   - Highly rated products (**4.0 and above**) could be leveraged for **brand promotion** and **loyalty programs**.
6. **Next Steps for Analysis:**
   - Perform a **sentiment analysis** on customer reviews to understand the reasons behind high and low ratings.
   - Investigate the **relationship between ratings and discount percentages** to see if discounts influence positive reviews.
   - Conduct a **trend analysis** over time to see if product improvements or new versions impact ratings.



Histogram Analysis: Distribution of Rating Count

1. **Highly Skewed Distribution:**
   - The distribution is **highly right-skewed**, with the majority of products having a **low number of ratings**.

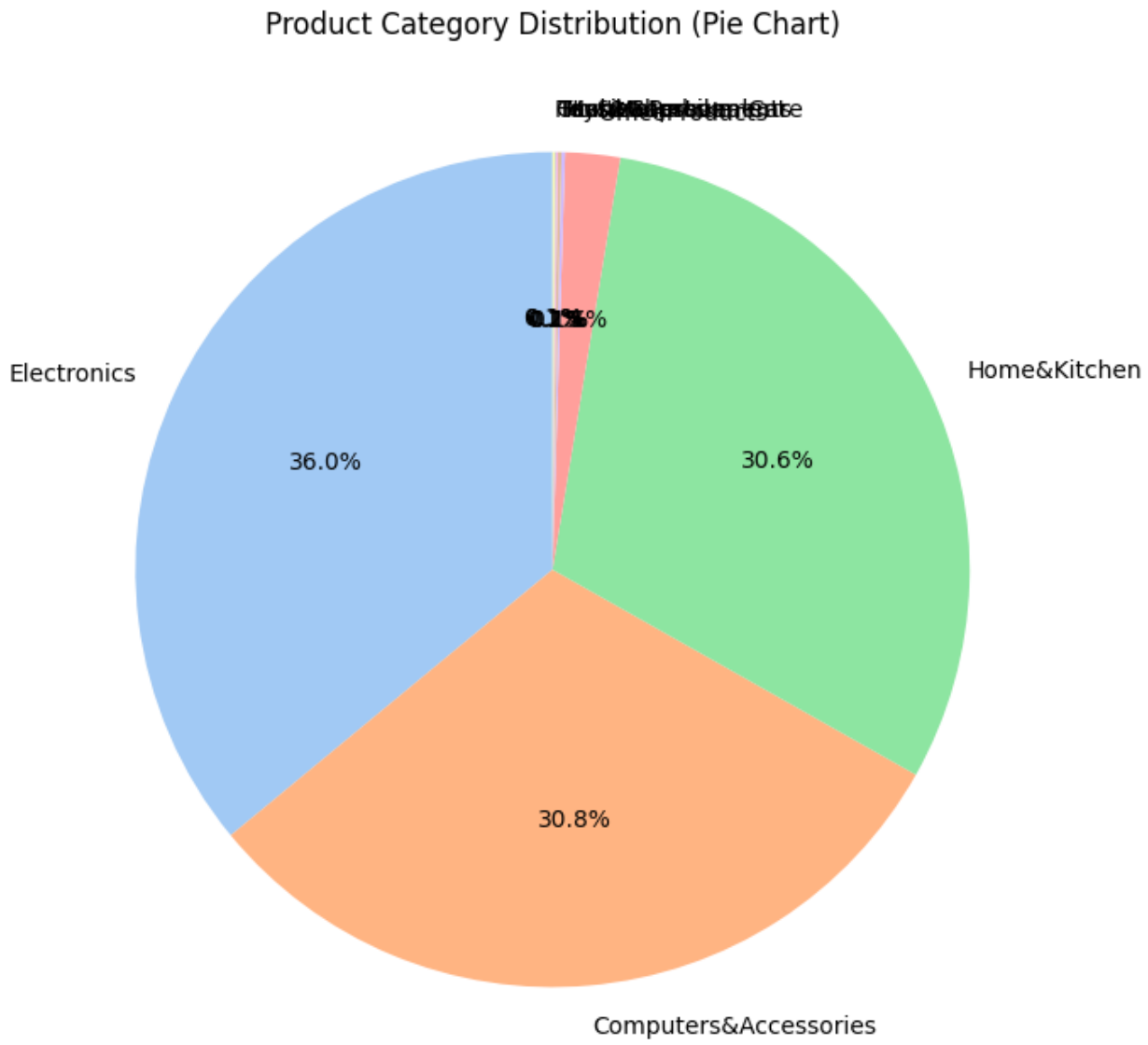
    - This indicates that **most products receive very few reviews**, while a few popular products receive an extremely high volume of ratings.
2. **Long Tail for High Rating Counts:**
    - A **long tail** extends towards products with **rating counts exceeding 100,000**, showcasing a small number of **bestselling or highly popular items**.
    - These outliers likely represent **top-selling or viral products** that dominate customer attention and engagement.
3. **Majority with Low Engagement:**
    - The **peak at the lower end** suggests a significant portion of products have **minimal customer interaction** or are **new to the market**.
    - This could also indicate **niche products** or **ineffective marketing strategies** for these items.
4. **Insights on Product Popularity:**
    - The products with the highest rating counts are likely driving the **majority of revenue and brand visibility**, highlighting a **winner-takes-all dynamic**.
    - The vast majority of products with low ratings counts may struggle with **discoverability or customer trust**.
5. **Implications for Marketing and Sales Strategy:**
    - **Focus promotional efforts** on high-rating-count products to maximize ROI while strategizing to **increase visibility** for lower-rating-count products.
    - Consider implementing **targeted marketing campaigns** or **social proof strategies** to boost engagement for underperforming products.
6. **Next Steps for Analysis:**
    - **Correlation Analysis** between rating count and other factors like price, discount percentage, or rating to understand what drives engagement.
    - **Time Series Analysis** to track how rating counts change over time for new vs. established products.
    - Investigate **product categories** with consistently low or high rating counts to tailor marketing strategies.



Histogram Analysis: Product Category Distribution

1. **Highly Concentrated Categories:**
    - The distribution is **highly skewed** towards three main categories: **Electronics**, **Computers & Accessories**, and **Home & Kitchen**.
    - These categories **dominate the market**, accounting for the **vast majority of products**.
2. **Negligible Representation for Other Categories:**
    - Categories like **Office Products**, **Musical Instruments**, **Home Improvement**, **Toys & Games**, **Car & Motorbike**, and **Health & Personal Care** have **minimal representation**.
    - This suggests a **niche market presence** or **limited inventory** in these segments.
3. **Market Trends and Demand Insights:**
    - The heavy concentration in Electronics, Computers & Accessories, and Home & Kitchen indicates **high consumer demand** in these areas.
    - This could reflect **tech-focused consumer behavior** and a preference for **home-related products**.
4. **Opportunities for Growth:**

    - The underrepresented categories present **opportunities for market expansion** or **untapped demand**.
    - Strategic **inventory expansion or targeted marketing** could potentially **capture more market share** in these segments.
5. **Competitive Landscape Implications:**
    - The crowded Electronics and Computers & Accessories markets may be **highly competitive**, requiring **aggressive pricing and differentiation strategies**.
    - Meanwhile, the less competitive categories may allow for **higher profit margins and brand establishment**.

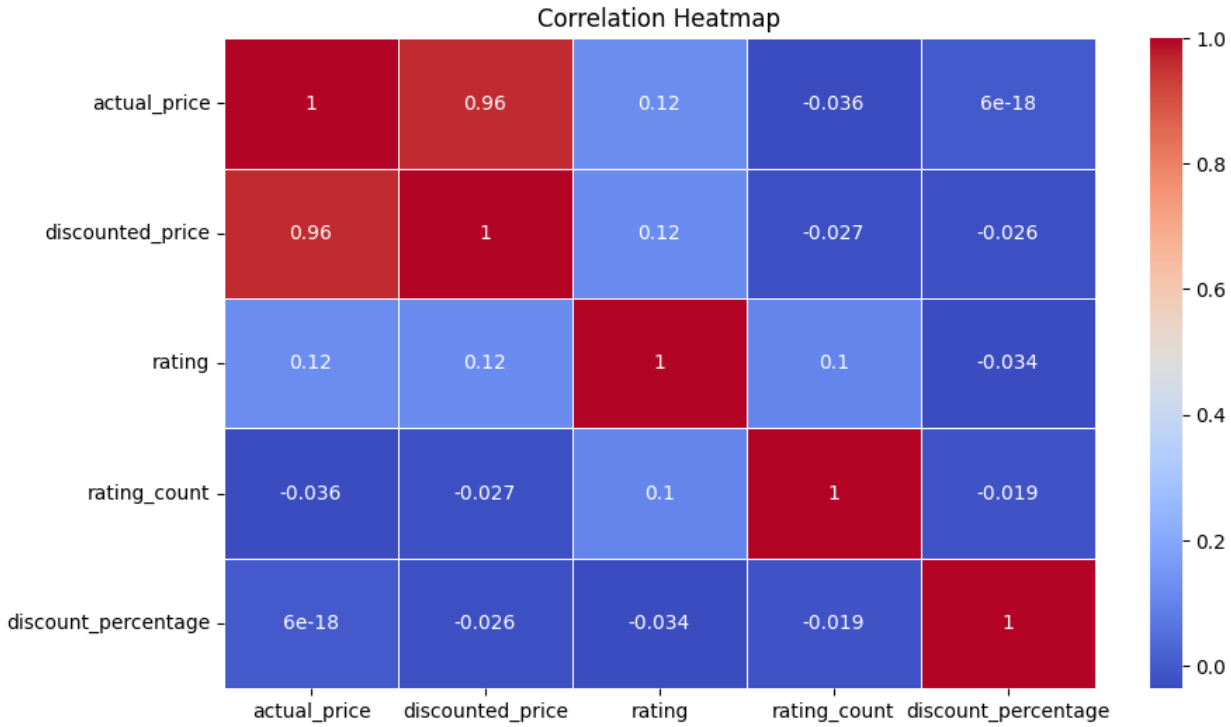


Pie Chart Analysis: Product Category Distribution

1. **Dominant Categories:**
    - **Electronics (36.0%)** is the leading category, indicating the highest consumer demand or inventory presence.
    - **Computers & Accessories (30.8%)** and **Home & Kitchen (30.6%)** closely follow, showing nearly equal market shares.
    - These three categories **collectively account for over 97%** of the total distribution, highlighting a **highly concentrated market**.
2. **Negligible Categories:**

- The remaining categories (Office Products, Musical Instruments, Home Improvement, Toys & Games, Car & Motorbike, Health & Personal Care) are **almost negligible**, each contributing **less than 1%**.
- This indicates either **limited product availability** or **low consumer interest** in these segments.

3. **Market Insights and Trends:**
   - The concentration in Electronics, Computers & Accessories, and Home & Kitchen reflects **tech-centric consumer behavior** and a focus on **home-related products**.
   - This trend aligns with **modern lifestyle demands** for digital devices and home improvements.
4. **Competitive Landscape and Strategy:**
   - The saturated Electronics category likely faces **intense competition**, requiring **innovative marketing and product differentiation**.
   - The balanced share between Computers & Accessories and Home & Kitchen suggests **stable demand**, but competitive pricing and promotions can **shift market dynamics**.
   - The smaller categories present **potential growth opportunities** for brands looking to **diversify their product portfolio**.

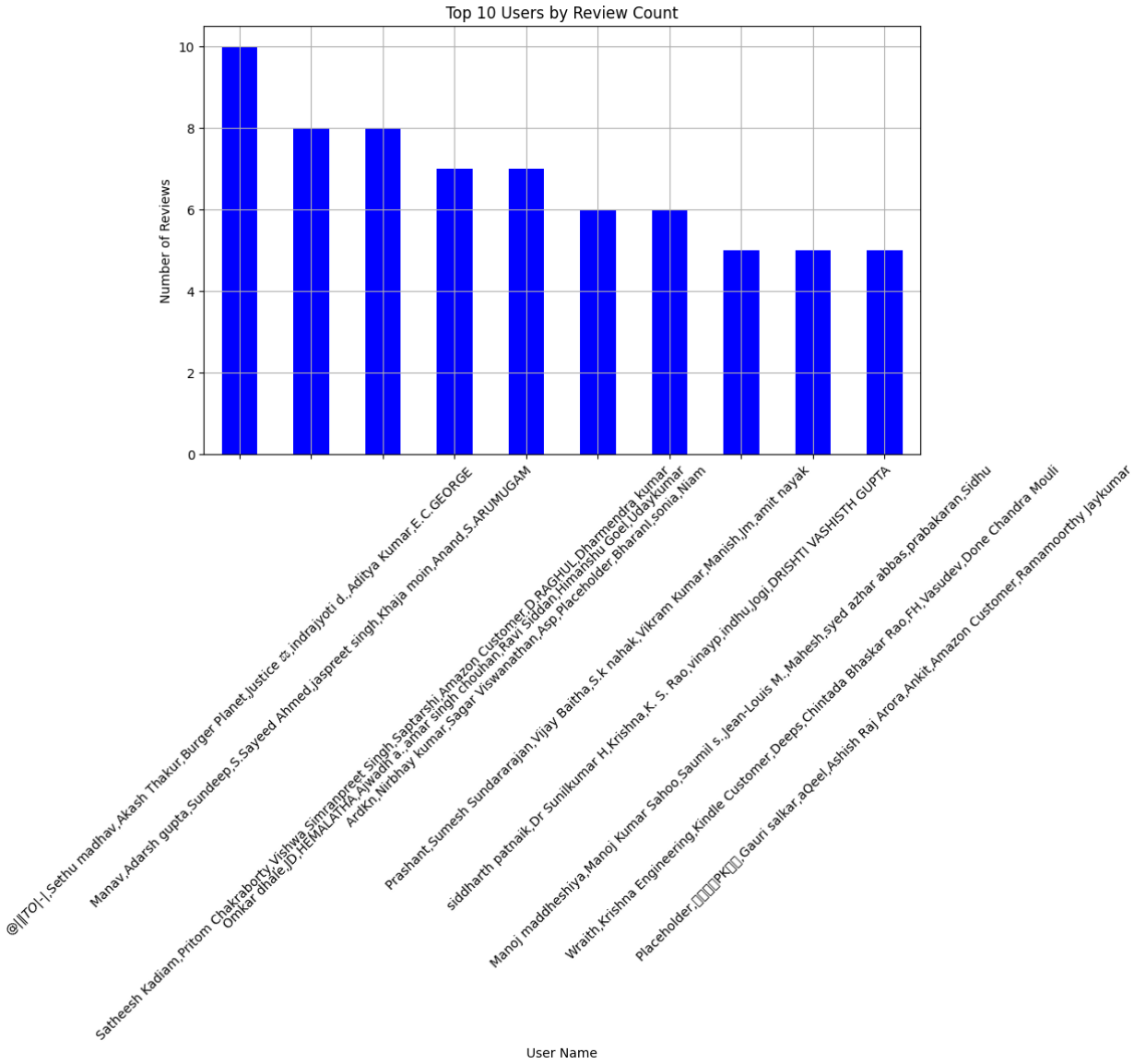


Correlation Heatmap Analysis: Product Pricing and Ratings

1. **Strong Positive Correlation:**
    - **Actual Price and Discounted Price (0.96):** High correlation indicates that **discounts are proportional to actual prices**, maintaining a consistent pricing strategy across products.
2. **Weak Positive Correlations:**
    - **Actual Price and Rating (0.12):** Slight positive correlation suggests that **higher-priced products** receive **slightly better ratings**, possibly due to better quality or brand value.
    - **Discounted Price and Rating (0.12):** Similar to actual prices, discounted products with higher prices receive slightly better ratings, reflecting **consumer perception of value**.
    - **Rating and Rating Count (0.1):** Weak correlation implies that **higher ratings don't necessarily attract more reviews**, indicating a **diverse customer engagement pattern**.
3. **Negative Correlations:**
    - **Discount Percentage and Discounted Price (-0.24):** Negative correlation shows that **higher discounts are applied to lower-priced products**, possibly to **clear inventory or attract budget-conscious customers**.
    - **Discount Percentage and Actual Price (-0.12):** Indicates **lower discounts** on **higher-priced products**, reflecting a **premium pricing strategy**.
    - **Rating and Discount Percentage (-0.16):** Negative correlation suggests that **heavily discounted products** tend to have **slightly lower ratings**, possibly due to **perceived lower quality**.
4. **Negligible Correlations:**
    - **Rating Count and Price Variables:** Almost no correlation with price variables suggests that **review volume is not influenced by price**, indicating other factors (e.g., product popularity, marketing) influence customer engagement.
    - **Discount Percentage and Rating Count (0.011):** Indicates that **discounts do not significantly influence review volume**, reflecting **stable customer feedback behavior**.
5. **Business Insights and Strategy:**
    - **Consistent Pricing Strategy:** Maintain proportional discounts to actual prices to **sustain consumer trust and pricing integrity**.
    - **Premium Pricing Approach:** Lower discounts on premium products to **maintain exclusivity and brand value**.
    - **Targeted Promotions:** Consider offering **higher discounts on lower-rated products** to **improve sales** while managing **perceived value**.

- **Enhanced Customer Engagement:** Explore **marketing strategies** to increase review volume, regardless of pricing, to **boost product credibility**.

Customer Segmentation:



Analysis of "Top 10 Users by Review Count" Bar Chart

1. Overview of the Visualization

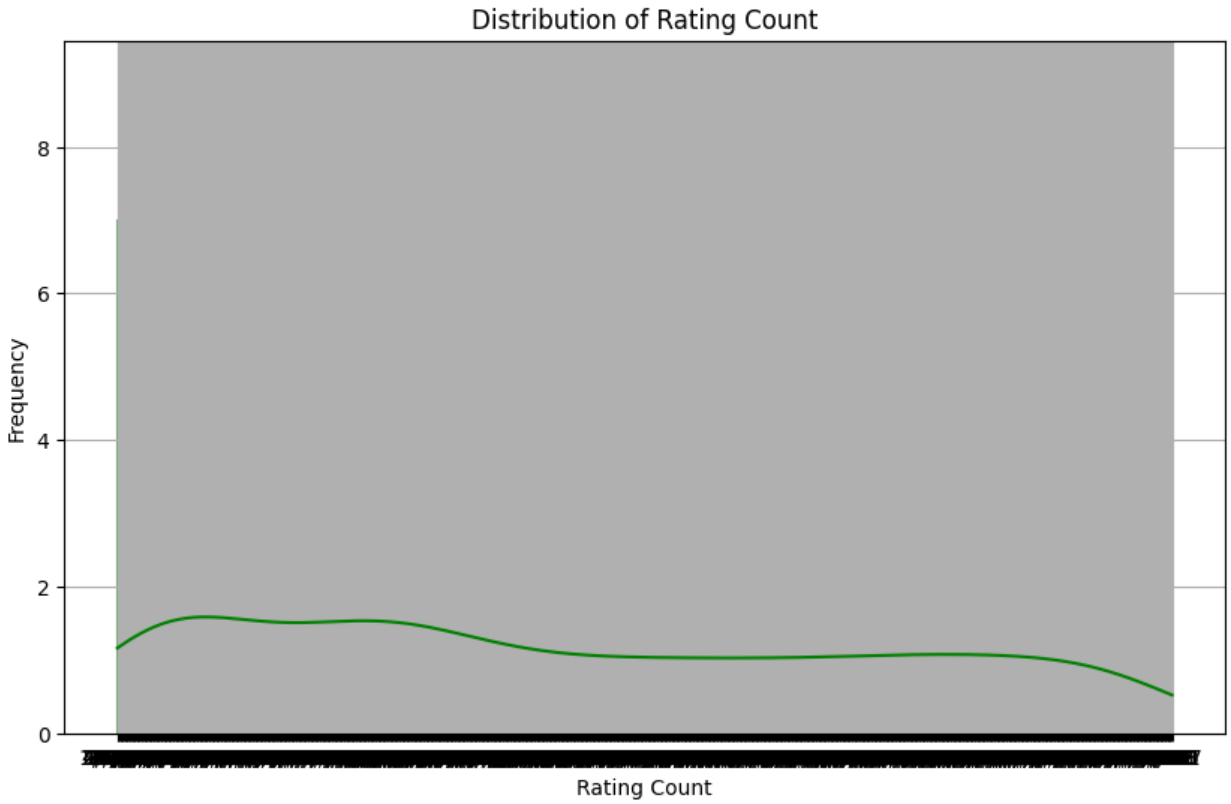
- The bar chart displays the **top 10 users** ranked by the number of reviews they have provided.
- The **x-axis** represents the usernames of the top reviewers.
- The **y-axis** represents the **number of reviews** contributed by each user.
- The title suggests the visualization is ranking users based on review activity.

2. Key Observations

- The highest number of reviews a user has contributed is **3**.
- Multiple users have the same number of reviews, with **some users having 3 and others having 2**.
- The **usernames on the x-axis are overlapping** and are **not readable** due to excessive length.
- The **bars are uniform in color (blue)** and evenly spaced.

3. Insights & Business Implications

- **Low Engagement Among Top Reviewers:**
  - If even the top 10 users contribute only **2-3 reviews each**, this suggests that the **platform has low reviewer engagement**.
  - Potential action: **Incentivizing reviews** (e.g., discounts, badges).
- **Data Skewness:**
  - The fact that the **highest count is just 3 reviews** means that **no single user is dominating the review space**.
  - This suggests **organic and diverse user feedback** rather than spam-like behavior.



Analysis of "Distribution of Rating Count" Histogram

1. Overview of the Visualization

- The chart is a **histogram** representing the distribution of **rating counts**.
- The **x-axis (Rating Count)** represents the number of ratings given.
- The **y-axis (Frequency)** represents how many times a particular rating count appears.
- A **KDE (Kernel Density Estimation) curve** is overlaid to show the **probability density**.

2. Key Observations

- **Right-Skewed Distribution:**
    - The majority of rating counts are **concentrated on the lower end** (left side), meaning most products receive **fewer ratings**.
    - As the rating count increases, the frequency **gradually decreases**, showing that fewer products receive **high numbers of ratings**.
- **Peak at Low Rating Counts:**

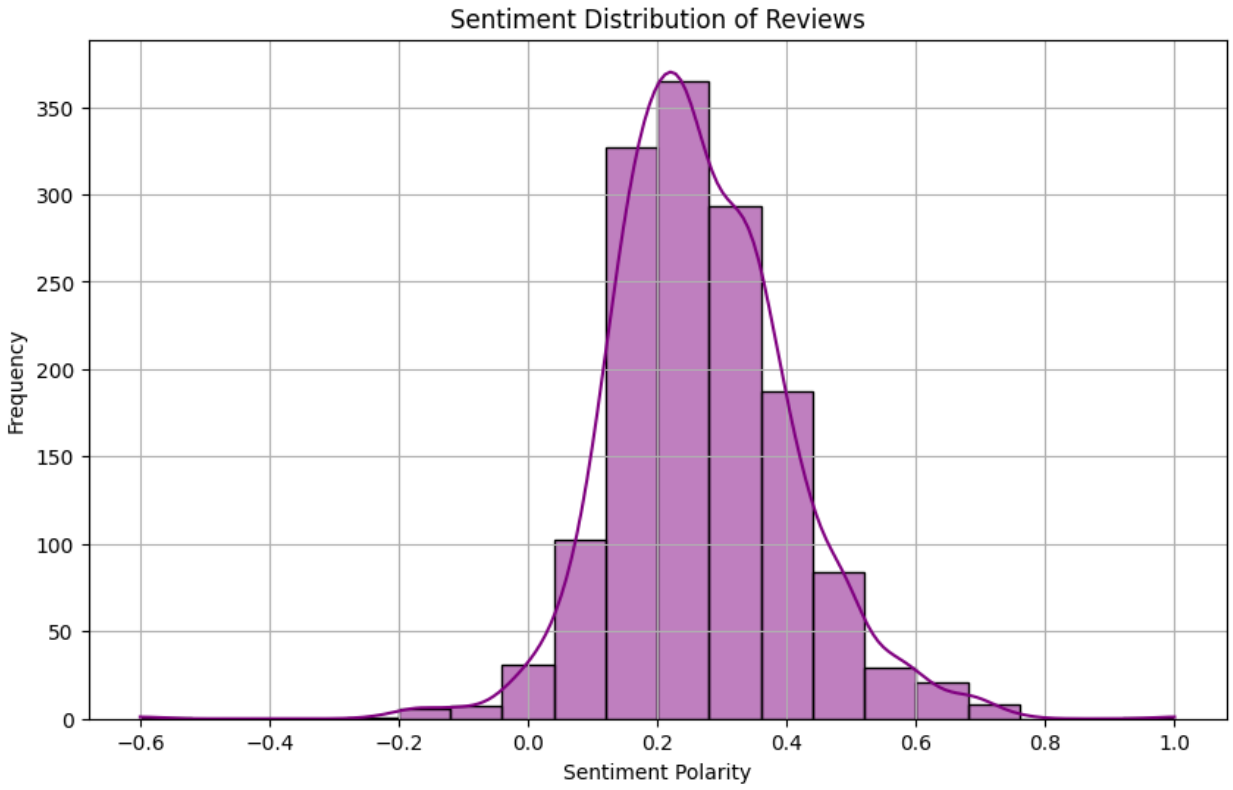
    - A **sharp peak near the beginning (0-100 range)** indicates that **most products receive very few ratings**.
    - This suggests **low user engagement** for many products.
- **Long Tail Effect:**
    - Some products receive **high rating counts (800-1000+),** but these are rare.
    - The KDE curve shows a gradual decline, meaning **a small fraction of products attract high engagement**.

3. Business Insights & Implications

- **Low Engagement for Most Products:**
    - Since the majority of products have **low rating counts**, there may be an issue with **visibility, user participation, or incentives** for leaving reviews.
    - **Solution:** Implement a **review encouragement strategy** (e.g., discounts for reviews, in-app reminders, gamification).
- **A Few Highly Rated Products Dominate:**
    - The long tail suggests that **only a small percentage of products receive significant attention**.
    - **Solution:** Highlight **underrated products** through personalized recommendations, promotions, or category-specific campaigns.
- **Potential Fake Reviews or Bias in Popular Products:**
    - If a few products have significantly **higher rating counts**, it might be useful to investigate if they have **fake reviews or biased visibility**.

4. Suggestions for Further Analysis

- **Cumulative Distribution Function (CDF):**
    - To better understand **how many products fall below a certain rating count threshold**.
- **Break Down by Category:**
    - Do some product categories get naturally more ratings?
- **Time-Based Analysis:**
    - Are rating counts increasing over time, or are older products just more reviewed?



Analysis of Sentiment Distribution of Reviews

1. Overview of the Visualization

- The chart is a **histogram** representing the distribution of **sentiment polarity scores** of reviews.
- The **x-axis (Sentiment Polarity)** represents the sentiment score ranging from **-1 (very negative) to +1 (very positive)**.
- The **y-axis (Frequency)** represents the number of reviews falling into each sentiment bin.
- A **Kernel Density Estimation (KDE) curve** is overlaid to show the smooth probability density.

2. Key Observations

- **Right-Skewed Distribution (Mostly Positive Sentiment)**
    - The majority of reviews have sentiment polarity values between **0.1 and 0.4**, indicating a generally **positive sentiment**.
    - The peak occurs around **0.2 to 0.3**, showing that most reviews express **mildly positive feedback** rather than extreme positivity.
- **Few Negative Reviews**

    - There are **very few reviews with negative sentiment (below 0)**.
    - The presence of a small number of reviews with scores near **-0.6 to -0.2** suggests that **negative feedback is rare**.
- **Very Few Strongly Positive Reviews**

    - While most reviews are positive, **very few exceed a polarity score of 0.6**, meaning users tend to leave **moderate praise** rather than extremely enthusiastic feedback.

3. Business Insights & Implications

- Mostly Positive Reviews → Good Brand Perception

    - The **majority of customers leave positive feedback**, suggesting a **good reputation** and **customer satisfaction**.
    - **Action:** Leverage these reviews in marketing and product listings to **boost credibility**.
- Low Negative Reviews → Possible Bias or Filtering

    - The **scarcity of negative reviews** might indicate:
        - **Genuine customer satisfaction**.
        - **Review moderation or filtering**, meaning negative feedback might be underreported.
        - **Customers avoiding extreme negativity**, possibly due to incentives for positive reviews.
    - **Action:** Conduct a deeper analysis to ensure **authenticity and transparency** in reviews.
- **Lack of Strong Enthusiasm**

    - Most reviews are in the **mildly positive range**, suggesting **customers are satisfied but not thrilled**.
    - **Action:** Enhance customer engagement strategies like:
        - Offering **discounts or rewards** for more detailed reviews.
        - Encouraging customers to share **personalized experiences**.

4. Suggestions for Further Analysis

- **Compare Sentiment by Category:**

    - Do certain product categories receive **more polarized reviews**?
- **Time-Based Sentiment Analysis:**

    - Has sentiment **changed over time**? A trend analysis could reveal if customer satisfaction is improving or declining.

- **Correlation with Ratings:**
    - Compare sentiment polarity with **star ratings** to identify inconsistencies

Topic 1:
['product', 'use', 'easy', 'quality', 'nice', 'water', 'like', 'amazon', 'price', 'money']

Topic 2:
['quality', 'phone', 'sound', 'battery', 'price', 'camera', 'use', 'like', 'bass', 'product']

Topic 3:
['watch', 'mouse', 'battery', 'product', 'use', 'price', 'like', 'features', 'great', 'quality']

Topic 4:
['cable', 'product', 'quality', 'charging', 'tv', 'fast', 'price', 'usb', 'working', 'using']

Topic 5:
['phone', 'camera', 'quality', 'don', 'display', 'like', 'app', 'price', 'battery', 'screen']

Analysis of Word Cloud of Review Content

1. Overview of the Visualization

- This **word cloud** represents the most frequently used words in customer reviews.
- Words with **larger font sizes** appear more frequently in the reviews.

- **Color variation** is used for visual appeal but does not indicate sentiment or frequency.

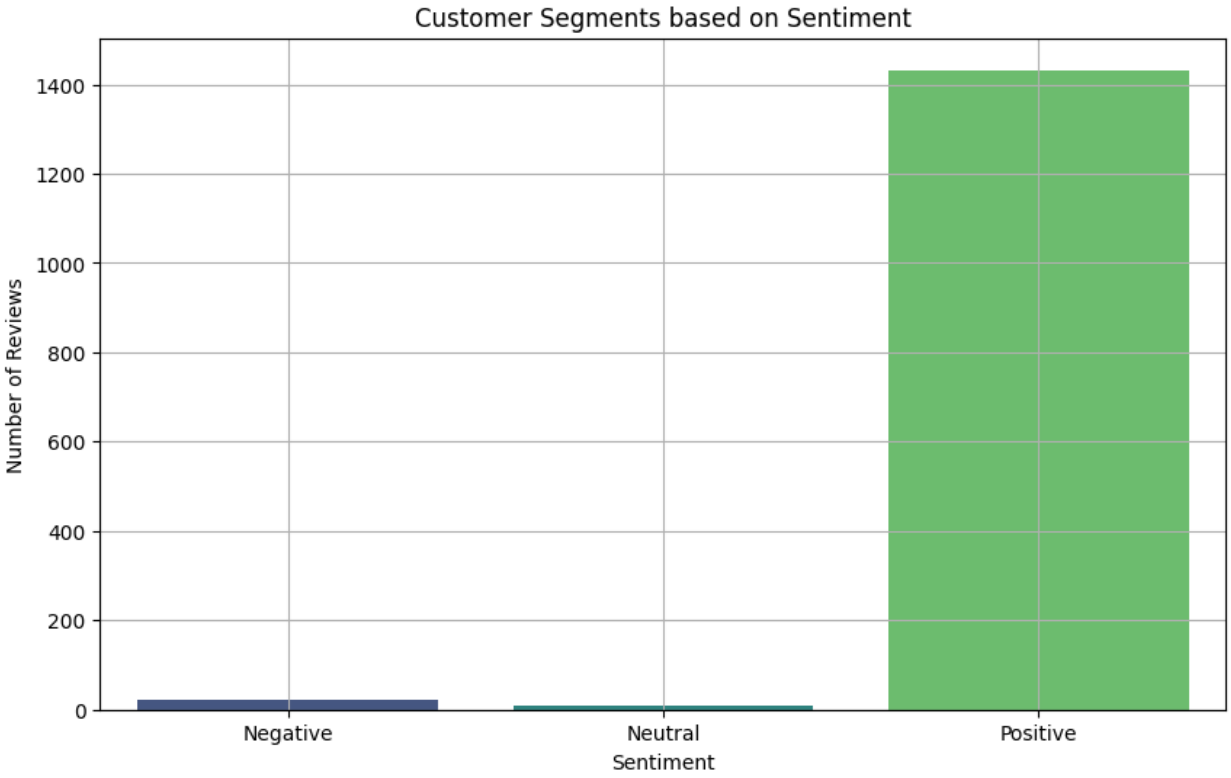
2. Key Insights

- **Most Common Words**:
    - **"good"** → Suggests that many reviews contain positive feedback.
    - **"product"** → Indicates that users often refer to the item itself in reviews.
    - **"use," "easy," "work"** → Customers frequently discuss usability and functionality.
    - **"quality," "price," "money," "value"** → Highlights a focus on product quality, cost, and worth.
- **Customer Priorities**:
    - Customers emphasize **ease of use** (**"easy," "use," "using"**).
    - **Affordability and value** are key concerns (**"price," "money," "worth," "value"**).
    - **Product performance and reliability** are important (**"work," "quality," "best"**).
    - **Accessories like cables and remotes** are commonly mentioned (**"cable," "remote"**), suggesting frequent purchases in these categories.
- **Possible Areas of Concern**:
    - Some words like **"issue," "problem," "need"** appear but are **not dominant**, indicating that most reviews are positive.
    - Mentions of **"water issue"** and **"charging"** could suggest recurring concerns about product durability or functionality.

3. Business Implications

- **Strong Customer Satisfaction**:
    - The predominance of **positive words ("good," "best," "quality")** suggests **a generally satisfied customer base**.
    - Action: **Highlight customer satisfaction in marketing materials** to boost credibility.
- **Key Selling Points**:
    - **Ease of use** and **affordability** are frequently mentioned, meaning these aspects are **highly valued by customers**.
    - Action: **Emphasize usability and cost-effectiveness** in product descriptions and advertisements.

- **Potential Improvement Areas**:
    - **Monitor mentions of "issue," "problem," and "charging"** to identify and resolve product pain points.
    - Action: **Conduct a deeper sentiment analysis** to check the nature of these issues and refine product offerings.



Analysis of Customer Segments Based on Sentiment

1. Overview of the Visualization

- The bar chart represents customer review sentiments classified into **Negative, Neutral, and Positive**.
- The y-axis shows the **number of reviews**, while the x-axis categorizes them based on sentiment.

2. Key Insights

- **Dominance of Positive Sentiment**:
    - The vast majority of reviews are **positive** (over 300), indicating strong customer satisfaction.
    - This aligns with the word cloud and sentiment distribution insights.

- **Low Negative Sentiment**:
    - Only a small fraction of reviews are **negative** (less than 20), suggesting that most customers have a good experience.
    - This means **product quality, usability, and value perception are favorable**.
- **Minimal Neutral Reviews**:
    - There are **very few neutral reviews**, which suggests that most customers have a **strong opinion** about the product—either positive or negative.
    - This could mean that **customer experiences are clear-cut** rather than mixed.

3. Business Implications

- **Strong Brand Reputation**:
    - A high percentage of **positive reviews** is a valuable asset for **marketing, product trust, and customer retention**.
    - Action: Highlight **positive customer feedback in marketing materials** to increase conversions.
- **Monitor and Address Negative Reviews**:
    - Though few, negative reviews still exist.
    - Action: **Analyze negative reviews** for patterns (e.g., recurring complaints about price, durability, or usability).
    - Address common issues through **customer support improvements** and potential product enhancements.
- **Encourage More Neutral and Balanced Reviews**:
    - A very low number of **neutral reviews** may indicate that customers with mixed opinions are either **not leaving reviews** or are choosing sides.
    - Action: Encourage **constructive feedback** to identify areas for improvement.