**Data Description**

* product\_id - Product ID
* product\_name - Name of the Product
* category - Category of the Product
* discounted\_price - Discounted Price of the Product
* actual\_price - Actual Price of the Product
* discount\_percentage - Percentage of Discount for the Product
* rating - Rating of the Product
* rating\_count - Number of people who voted for the Amazon rating
* about\_product - Description About the Product
* user\_id - ID of the user who wrote review for the Product
* user\_name - Name of the user who wrote review for the Product
* review\_id - ID of the user review
* review\_title - Short review
* review\_content - Long review
* img\_link - Image Link of the Product
* product\_link - Official Website Link of the Product

**Insights from the Data:  
Key insights:**

* there are 1465 unique reviews, which consist of 1351 unique products, and 1194 unique users
* the Category feature could be more useful if we break it down into subcategories using one-hot-encoding
* all numerical variables have object data types, which we will convert into float for EDA and model building.
* Otherwise, the data is mostly clean with very few missing data, so we will consider dropping the null entries.

**Data Cleaning & Feature Engineering**

1. convert numerical features to the correct type
2. drop null values (since it's only a small amount
3. Add length of the review content
4. Expand feature space for **category** (only the first two for each reviewed product, as we reason that the first two are usually the more general categories.)

**Approach to feature Engineering:**

Specifically, we tackle the 'category' column where each entry is a string containing multiple subcategories separated by a pipe symbol '|'. The process starts by identifying and extracting the first two subcategories from each entry to capture the primary and secondary category information.

Once these unique subcategories are identified, we will generate new features (columns) in the dataset for each subcategory. These features are binary, indicating the presence (1) or absence (0) of each subcategory within the first two positions of the original 'category' entry for each record. This binary encoding allows for the categorical data to be easily used in various later data analysis techniques, such as clustering or machine learning models, where numerical input is required.

The outcome is a data frame with expanded features, enabling a more granular analysis of the data based on categorized attributes. This method is particularly useful in customer segmentation, product categorization, and other applications where categorical distinctions are important.

**Exploratory Data Analysis (EDA):**

**General Observations:**

**From Histograms of Numerical Features**

* The right skew in the price histograms (both discounted and actual) suggests that while most products are relatively inexpensive, there are a few products with very high prices.
* The rating count and the discount percentage histograms provide insights into customer engagement and pricing strategies. A high number of products with middle-range discount percentages might suggest a common discounting strategy, while the rating count reflects varying levels of customer interaction.
* The distribution of ratings shows customer satisfaction is generally high across the products, or that there is a bias in the dataset towards higher ratings (this can happen if customers who are dissatisfied are less likely to leave a review).

**From Correlation Matrix**

We don't seem to find anything significant in the correlation matrix. We do expect a strong positive correlation between the actual price and the discounted price. On the other hand, there is a very weak correlation between rating and price.

Interesting finding:

* a moderate positive correlation between price and review length. It means that as the price of a product goes up, reviewer tends to give longer reviews.

**From Bar Chart (top 10 product categories)**

Noted, we extract primary and secondary category information from each category string, It is possible that a product belongs to both Electronics ad Computer&Accessories. And it is also possible that one category (such as Computers&Accessories) is a sub-category of another category (such as Electronics).

* **Electronics** is the most reviewed product category, with just under 500 counts, making it the most popular or most frequently reviewed category among the dataset's users.
* **Computers & Accessories** follows closely behind, indicating high interest or purchase rates
* The third place is held by **Home & Kitchen**, which suggests a strong interest in products related to home improvement and kitchenware.
* **Accessories & Peripherals** and **Kitchen & Home Appliances** come next, with over 200 reviews each, pointing to a significant interest in these subcategories.
* The middle tier of the chart features categories like **Home Theater, TV & Video**, **Mobiles & Accessories**, and **Heating, Cooling & Air Quality**. These are likely important but less dominant categories in terms of review counts compared to the top categories.
* **Wearable Technology** and **Headphones, Earbuds & Accessories** have fewer reviews but are still significant enough to make it into the top 10, indicating a niche but notable market segment.
* This visualization helps stakeholders understand where the focus of customer feedback is and could inform inventory, marketing strategies, and customer support initiatives.
* The data suggests that electronics, computing, and home-related products are the primary interests within this dataset's scope. Firms in these sectors might have a larger customer base and could potentially capitalize on cross-selling opportunities among these closely related categories.

**Word Cloud Analysis:**

**Review content & Review title**

We use word clouds generated from "review content" and "review title" columns to provide a visual representation of the frequency of word usage within these textual data fields. In word clouds, the size of each word indicates its frequency or importance in the corpus from which the word cloud is generated.

Noted: we are focusing on products that belong to "electronics" category.

In the "Word Cloud for Review Contents":

* Words like "good", "product", "sound", "quality", "camera", "battery", "price", and "use" stand out as the most prominent. This suggests reviewers often mention the product's quality and features, particularly sound, camera quality, and battery life. Price appears to be a significant factor in the reviews, indicating that customers are mindful of the cost and its relation to the value they perceive from the product. The frequent mention of "use" may point to a focus on the usability and functionality of the products.

In the "Word Cloud for Review Titles":

* Like the content word cloud, terms like "good", "product", "quality", "price", and "sound" are quite prominent. However, there are additional words like "money", "worth", "great", and "best" that are more noticeable here. This could indicate that review titles tend to summarize overall satisfaction or value judgments, with a strong emphasis on the economic aspect of the purchase ("money", "worth"). Words like "great" and "best" could reflect positive product endorsements.

By examining both word clouds, quality and price are the most critical factors for reviewers, as these terms dominate both visualizations. This implies that consumers place a high value on the price-to-quality ratio when evaluating products. The presence of specific product features like "sound" and "camera" in the reviews suggests these are common points of discussion, indicating important deciding factors for potential buyers. **The positive terms like "good", "great", "best", and "amazing" suggest a generally positive sentiment in the reviews, which confirms our finding in the histogram, where most of rating are 4 - 5 (out of 5)**

These word clouds can be incredibly useful for product manufacturers and sellers, providing insights into what aspects of a product are most important to customers, which features are most often praised or criticized, and what factors are most influential in a purchase decision.

**Product Name**

**Technical Specifications:**

Words like "RAM," "128GB," "Storage," "4K," "HD," "Bluetooth," "LED," "USB," and "HDMI" are quite prominent. This suggests that technical specifications are a significant focus in the product descriptions, which is typical for electronic items where specifications are important for consumer decisions.

**Product Categories:**

Terms like "Smart," "Watch," "TV," "Phone," "Speaker," "Tablet," "iPhone," and "iPad" are indicative of the various categories of electronics that are being offered. The prominence of "Smart" and "Pro" suggests a trend towards high-end or professional-grade products.

**Connectivity and Portability:**

"Wireless," "Portable," "Bluetooth," and "Cable" imply that connectivity and ease of transport are important selling points. "Wireless" and "Bluetooth" likely refer to the ability of devices to connect without cords, which is a valued

feature for convenience and mobility.

**Display and Visual Features:**

"4K," "HD," "Display," "AMOLED," "LCD," and "Screen" are prominently featured, highlighting that display quality is a key aspect of product descriptions. High-definition visuals are a crucial selling point for TVs, smartphones, and tablets.

**Performance and Capacity:**

High-frequency words like "Fast," "Speed," "Charging," "128GB," and "RAM" emphasize performance metrics and storage capacity. "Fast Charging" is particularly notable, pointing towards the consumer's desire for quick recharge times in their devices.

**Audio and Health Features:**

"Speaker," "Earphone," and "Heart Rate" suggest a focus on audio products and health monitoring devices, indicating a market interest in personal audio and fitness tracking features.

**Design and Quality:**

"Ultra," "Smart," "Compact," "Durable," and "Quality" indicate that the physical design and build quality are important characteristics detailed in product descriptions. The repeated use of "Smart" also points to a market trend towards smart technology integration.

**Brand and Model References:**

Brand names like "Samsung" and "iPhone" and model indicators like "Series" suggest that brand recognition and specific model details are commonly highlighted, which is important for consumers who are brand-conscious or looking for the latest models.

**Functionality and Use Cases:**

"Remote," "Calling," "Compatible," "Sports," and "Mode" reflect the various functionalities and use cases of the products. It implies that products are being marketed for their utility in specific contexts, such as sports or compatible with other devices.

Overall, the word cloud indicates that the electronic products on Amazon are being marketed with an emphasis on **technical specifications, connectivity, display quality, and brand recognition**. The prominence of certain terms reflects what sellers believe are the key selling points and what consumers may prioritize when browsing electronic products. This data can be instrumental for sellers to understand market trends, consumer preferences, and competitive product features.

**Clustering Analysis:  
Clustering - Hierarchical**

**1. Goal: Given the product's detailed description, can we find the hierarchy of its categorization? In other words, what product category should be grouped together given their similarities?**

* Improved Product Discovery: By accurately categorizing products into a hierarchical structure, businesses can enhance the search and discovery process on their platforms. This makes it easier for customers to find what they're looking for, improving the overall shopping experience.
* Targeted Marketing: Understanding product hierarchies enables more precise segmentation for marketing campaigns. Products can be grouped into relevant categories and subcategories, allowing for tailored marketing messages that are more likely to resonate with specific customer segments.
* Inventory Management: A clear view of product categorization helps in better inventory management. Businesses can identify which categories are most popular or underrepresented, informing stock decisions and helping to optimize inventory levels.

Here’s the breakdown of the process:

* **Feature Extraction**: You start by converting the text descriptions of products into a numerical format that can be processed by clustering algorithms. This is done using the TF-IDF Vectorizer, which transforms the about\_product text into a matrix of TF-IDF features. TF-IDF highlights words that are more relevant to each product, thus preparing the data for further analysis.
* **Hierarchical Clustering**: With the features extracted, we then apply hierarchical clustering to the TF-IDF matrix. This method groups products based on the similarity of their descriptions, forming a hierarchy from the most distinct groups to a single cluster containing all products. The Ward method is used here, which minimizes the variance within each cluster.
* **Cluster Assignment**: After forming the hierarchical structure, you decide on a number of clusters to focus on (40 in this case) and assign each product to one of these clusters. This step is crucial for simplifying the complex hierarchy into a more manageable number of groups.
* **Identifying Common Categories**: For each of the 40 clusters, you analyze the categories of the products within them to find the most common category. This involves looking at the last part of each product's category path (indicating the most specific category) and identifying which category is most frequent in each cluster.
* **Visualization**: Finally, you visualize the hierarchical clustering results with a dendrogram, but with a twist. Instead of labeling each leaf with product names or IDs, you label them with the most common category found in each cluster. This provides a clear visual representation of how products are grouped based on their descriptions and the dominant category in each group.

**Cluster Hierarchy**: We can see how each cluster is grouped together. For instance, examining the leftmost tree nodes, clusters with dominant categories such as Mice and Wireless USB Adapters merge first, then they join with USB Cables, and eventually with Smart Televisions. This indicates that, based on the product descriptions, the clustering algorithm suggests Mice and Wireless USB Adapters products are similar; USB Cables are akin to both Mice and Wireless USB Adapters products, and so on.

**2. Similar Goal: Given the product review contents/title, can we find the hierarchy of its categorization? In other words, what product (categories) have similar review content/title?**

Customer Insight and Experience: Reviews are rich in customer insights and reflect actual user experiences and perceptions. By clustering reviews into categories, a business can better understand how customers are engaging with products and what aspects are most important to them. This can inform improvements in user experience and product design.

**Summary**

In the three use cases (using Product Description, Review Content, and Review Title), it is more logical to perform hierarchical clustering on the TF-IDF matrix derived from the product description data. This is because the review content may only serve as an indirect reflection of the product itself; some products can have similar reviews yet be very different in terms of categories.

**K-means Clustering**

For review content and titles, it would be more advantageous to conduct review segmentation. We could apply K-means clustering to categorize user reviews into distinct groups. This approach would aid in understanding user preferences and their correlation with rating scores and sentiment analysis. Additionally, such segmentation can uncover patterns in customer feedback that are not immediately apparent. It can also highlight specific features or issues that are common within groups of reviews, providing actionable insights for product refinement and customer service improvement. By analyzing these segments, businesses can also tailor their communication strategies to address the concerns and preferences of different customer segments more effectively.

* The plot displays a wide dispersion of review lengths across different rating counts.
* Different colors represent the 10 different clusters/categories identified by the K-means algorithm. The clusters appear to be distributed across all ranges of rating counts and review lengths, suggesting that the clusters may represent different characteristics beyond these two features (**which is captured by the review content in the tf-idf**) .
* There is a weak positive correlation that can be deduced between review length and rating count from this plot alone. It does appear that longer reviews correlate with a higher up-vote.

**RFM Analysis**

|  | General  Segment | Average  Frequency | Min  Frequency | Max  Frequency | Average  Monetary | Min  Monetary | Max  Monetary | Customer  Count |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | Gold | 1.932270916 | 1 | 10 | 19805.14606 | 1995 | 241600 | 251 |
| 1 | Silver | 1.157303371 | 1 | 3 | 6964.85764 | 999 | 75990 | 178 |
| 2 | Bronze | 1.011811024 | 1 | 2 | 2311.7821 | 39 | 59900 | 762 |

**Gold Segment**

• Average Frequency: Customers in this segment make purchases more often than those in other segments, with an average frequency of about 1.93. This suggests a high level of engagement with your products or services.

• Monetary Values: They also spend significantly more, with an average monetary value of 19,805, which is much higher than the other segments. The range of spending is from 1,995 to 241,600, indicating a diverse group of high spenders that may include both consistently high-value purchasers as well as those making occasional large purchases.

• Customer Count: The Gold segment is substantial in size, with 251 customers. Considering their high frequency and monetary value, these customers are likely the most valuable to your business and should be the primary focus for retention and loyalty programs.

**Silver Segment**

• Average Frequency: This segment shops less frequently, with an average frequency of about 1.16. This indicates occasional purchasing patterns.

• Monetary Values: Their spending is moderate, with an average monetary value of 6,964.85, ranging from 999 to 75,990. This suggests a mix of mid-tier customers who have the potential to be nurtured into more frequent and higher-value purchasers.

• Customer Count: There are 178 customers in the Silver segment. This group may include newer customers or those with specific needs. Tailored marketing campaigns could encourage them to increase their frequency and spending.

**Bronze Segment**

• Average Frequency: Customers in this segment have the lowest average frequency of about 1.01, implying that they typically make single or very few purchases.

• Monetary Values: Their average monetary value is the lowest at 2,311.78, with spending ranging from 39 to 59,900. Despite the low average spend, the wide range suggests some Bronze customers occasionally make significant purchases.

• Customer Count: This is the largest segment with 762 customers, but given their low engagement and spending, the focus should be on re-engaging them through special offers or by understanding their low activity through surveys and feedback.

**Implications for Marketing Strategy**

• For the Gold segment, the strategy should focus on exclusivity and recognition. Since they are frequent and high-value customers, personalized services, exclusive offers, and premium loyalty programs are effective. They are likely to respond well to upselling and cross-selling strategies.

• The Silver segment requires strategies that encourage them to purchase more frequently and potentially spend more per transaction. This could include offering them bundle deals, limited-time offers, or loyalty points that increase in value with the frequency of purchase.

• For the Bronze segment, the aim should be to understand their needs and reasons behind the infrequent and low-value purchases. Since this segment is quite large, there is a significant opportunity to move them up to higher segments. Initial engagement strategies such as first-time buyer offers or feedback surveys can help in increasing their activity.