# Best Value for Money Products on Amazon

Gabriela Todorova s2696177
Miglena Pavlova s2717972
Manojkumar Muthukumaran s3334325
Raja Ravi Varma Ramesh Babu s3441261
Akhila s3276562

## Introduction

In today's world, online platforms like Amazon have transformed shopping into a seamless and highly accessible experience, bringing a vast array of products—from electronics to books and beyond—directly to consumers. As this shift has taken hold, customer reviews and ratings have become a cornerstone of how people make purchasing decisions. These reviews not only shape opinions about which products to buy but also reveal important insights into how consumers feel about product quality, pricing, and overall value. Among all the factors influencing buyer decisions, value-for-money has emerged as one of the most significant. Yet, the relationship between customer ratings, product popularity, and real value-for-money often remains murky and underexplored. Better understanding this relationship is crucial—not just for shoppers trying to make informed choices but also for businesses aiming to meet customer needs while staying competitive.

Studying value-for-money carries implications far beyond individual purchasing decisions. For companies, understanding whether their products truly provide value to customers can refine pricing strategies, guide marketing efforts, and even shape product development. Products perceived as offering great value often gain consumer trust and loyalty, creating long-term benefits for brands. On the flip side, a disconnect between perceived and actual value can harm a company's reputation, spark negative reviews, and lead to lost sales. By analyzing patterns in customer reviews, pricing, and sales data, this research provides a structured way to evaluate value-for-money across different products and categories, helping to close the gap between consumer expectations and business strategies.

This work is particularly timely given the big scale and complexity of Amazon's marketplace. With millions of products and reviews, Amazon offers a unique and unparalleled dataset for identifying trends related to value-for-money. However, analyzing such large-scale data comes with its own challenges, from integrating diverse information to ensuring computational efficiency. To tackle these issues, this study employs PySpark, a powerful framework for distributed data processing. By using PySpark's ability to handle massive datasets, we aim to uncover actionable insights that have real-world applications for businesses, consumers, and

researchers alike. The outcomes of this study are going to benefit a wide range of stakeholders. For consumers, the findings can act as a roadmap to make smarter, value-driven purchases within their desired categories. Businesses, on the other hand, can use these insights to align their offerings more closely with what customers truly value. From a research standpoint, this work adds to the growing field of e-commerce analytics, demonstrating the practical potential of big data tools like PySpark to solve meaningful problems. By focusing on value-for-money, this research emphasizes the critical role customer perceptions play in shaping the future of online marketplaces and highlights the importance of bridging data with actionable insights.

# Related Work

Several studies have explored the relationship between customer reviews, product ratings, and value-for-money in online marketplaces. Research by Mudambi and Schuff (2010) highlights how review helpfulness plays a key role in shaping consumer perceptions, particularly for high-involvement products like electronics. Similarly, Chevalier and Mayzlin (2006) found that the volume and sentiment of online reviews significantly influence product sales, suggesting that customer feedback is a strong indicator of perceived value. Another study by Hu, Liu, and Zhang (2008) delves into the impact of star ratings, noting that while high ratings often correlate with purchase decisions, they do not always align with actual product quality or value-for-money. More recently, Luo (2024) examined pricing strategies on Amazon, analyzing how discounts impact product sales and customer ratings, demonstrating that perceived value is not solely dictated by price but also by the interaction between product reputation, review consistency, and competitive alternatives. Our research builds on these insights by not only analyzing value-for-money within Electronics and CD & Vinyl categories but also exploring the factors within high-rated reviews that influence consumer perceptions of value.
In our analysis, we explored several established formulas for ranking products based on value-for-money. The Bayesian Average Rating is a highly recommended method that smooths out anomalies by preventing obscure products with only a few 5-star ratings from ranking too high. The Wilson Score Interval offers a reliable way to rank products by incorporating uncertainty, making it useful for prioritizing items with a large number of positive reviews while accounting for mixed feedback. Additionally, Amazon’s heuristic-based approach provides a more realistic ranking by factoring in verified purchases, helpfulness votes, and review recency, ensuring that more credible and relevant reviews have a greater influence.

# Research Questions

Our study is built around four key questions that help us understand what makes a product truly worth its price. RQ1 looks at how products within the Electronics category compare in terms of value-for-money, helping to highlight which ones consistently offer the best bang for the buck. RQ2 takes this a step further, exploring how many highly rated products actually deliver good value, revealing any gaps between customer ratings and real-world worth. RQ3 dives into what

factors in high-rated reviews influence how customers perceive value-for-money in Electronics—whether it's durability, performance, or something else. Lastly, RQ4 zooms out to compare CD & Vinyl products with Electronics, identifying key similarities and differences in how value-for-money is perceived across these two very different categories.

*RQ1: How do products in the category “Electronics” compare in terms of value-for-money?*
*RQ2: What percentage of high-rated products are actually good value for money?*
*RQ3: What aspects mentioned in the high-rated reviews of the "Electronics" category most influence customers' perceptions of value-for-money?*
*RQ4: How do products in the "CD & Vinyl" category compare to those in the "Electronics" category in terms of value-for-money, and what similarities or striking differences can be observed?*

# Method

## Dataset

The dataset used in our research is a comprehensive collection of Amazon product information and customer reviews. It is divided into two main components: metadata and reviews. The metadata files contain detailed information about the products across various categories, such as Electronics, Books, Automotive, and more. Each metadata file includes essential attributes such as the *asin* (Amazon Standard Identification Number), which uniquely identifies the product, and the *title*, which provides a succinct name or description of the product. Additional fields include the *description*, offering a textual overview of the product's features and functionalities, and *categories*, which organizes the product into a hierarchical structure of its market placement. Furthermore, metadata often includes supplementary fields like *price*, specifying the listed value of the product, and *imUrl*, a direct link to an image representing the item. Advanced fields like *salesRank*, indicating the product's position in its category, and *related*, which identifies products that are often bought together, enhance the dataset's depth.

The reviews data complements the metadata by capturing user-generated content for each product, also organized by categories. Each review is associated with a product through the *asin* field, creating a direct linkage between the metadata and review data. The review records include critical fields such as *reviewerID*, a unique identifier for the user, and *reviewText*, the textual content where users share their opinions and experiences. Reviews also feature a numeric rating (*overall*) ranging from 1 to 5 stars, alongside a *summary* field that succinctly describes the review's essence. The *helpful* field records the number of users who found the review helpful, reflecting the community's interaction with the feedback. Time-related fields like *unixReviewTime* and *reviewTime* indicate when the reviews were written, helping analyze their distribution over time.

Table 1 below outlines the structure of the dataset, highlighting the main fields in both the metadata and review components.

| **Component** | **Field Name** | **Description** |
|---|---|---|
| Metadata | asin | Unique identifier for the product |
| | title | Name or description of the product |
| | description | Textual overview of the product features |
| | categories | Hierarchical categories in which the product is classified |
| | imUrl | URL link to the product image |
| | price | Listed price of the product |
| | salesRank | Sales rank of the product in its category. |
| | related | Related products categorized into *also_viewed*, *also_bought*, *bought_together* |
| Reviews | reviewerID | Unique identifier for the reviewer |
| | asin | Product identifier linking the review to its metadata |
| | reviewText | The full text of the user's review |
| | overall | Star rating (1-5) provided by the reviewer |
| | summary | Short summary of the review |
| | helpful | A tuple indicating how many users found the review helpful out of the total who evaluated it |
| | unixReviewTime | Unix timestamp indicating when the review was written. |
| | reviewTime | Human-readable date of the review (e.g., MM DD, YYYY). |

Table 1. Amazon dataset structure

# Preparation

While each of the approaches discussed in the *Related Work* section has its strengths, we ultimately decided on a Weighted Value-for-Money Ratio, which balances customer satisfaction, product popularity, and affordability. The formula we used is:

*Weighted Value-for-Money Ratio = (Average Rating × log(Number of Reviews + 1)) / Price*

This approach ensures that products with high ratings and strong review volumes are prioritized while also considering affordability, making it an effective metric for assessing real value-for-money across different product categories.

# Data Pre-Processing

To achieve meaningful insights, we implemented a structured approach to clean, organize, and integrate the data, making it suitable for computational analysis. This preprocessing phase ensured that our dataset was well-structured and ready for value-for-money calculations.
We started by extracting two datasets from the Amazon database. The first dataset contained product metadata, which included unique product identifiers, product titles, and prices. The second dataset consisted of customer reviews, capturing user-generated ratings and textual feedback, providing insights into customer satisfaction.
Before proceeding with calculations, we filtered and cleaned the data to ensure consistency and accuracy. Products with missing or zero price values were excluded from the dataset to prevent distortions in VFM calculations. Since some products had multiple reviews, we grouped the reviews by their unique identifiers and computed two key metrics: the average rating, representing an indicator of customer satisfaction, and the total number of reviews, which served as a measure of product popularity. The aggregated review data was then merged with the product metadata, creating a unified dataset that linked product details with customer feedback. This preprocessing step ensured that the dataset was structured and ready for the subsequent value-for-money analysis.

We included only records with a valid value-for-money score, ensuring that all relevant columns contributing to the metric had complete and accurate data. After applying these criteria, our final dataset for the Electronics category consisted of approximately 300,000 records. For finding the percentage of records that are actually good value for money from the top rated products, we filter the dataset based on ratings (>4.5). Then we create another dataframe with the filtered dataframe. This dataframe has all the records which fall in the top 50 percentile of VFM score. We then calculate the count number of records in each of these dataframes and calculate the ratio and percentage. To assess how customers perceive value-for-money (WVFM) in highly rated products, we conducted a structured analysis using Amazon reviews data. Our approach involved three key steps. First, we filtered the dataset to focus only on the top 50% of products ranked highest in value-for-money, ensuring that our analysis was centered on products that offer the best balance of price and performance. We extracted key attributes such as product ID (ASIN), review text, and overall rating, allowing us to concentrate on customer feedback specific to these high-ranked products. Next, we processed the review text by removing punctuation, converting it to lowercase, and tokenizing it into individual words to identify key aspects influencing customer perception of value-for-money. We focused on frequently mentioned attributes such as price, quality, features, design, value, performance, functionality, and durability, enabling us to determine which factors customers emphasized most in discussions

about top WVFM products. To measure the impact of these aspects on customer satisfaction, we categorized reviews into positive (4-5 stars) and negative (1-2 stars) groups and analyzed the frequency of each keyword within these categories. This step helped us uncover valuable insights, such as quality and performance being crucial in both positive and negative reviews, while price was frequently praised in positive reviews but less often criticized in negative ones. Additionally, durability concerns were more commonly associated with negative feedback, indicating that perceived lack of longevity significantly impacted customer ratings. By following this structured approach, we gained a data-driven understanding of the factors that matter most when customers evaluate a product's value-for-money. These insights can help businesses optimize product features, pricing strategies, and marketing efforts to enhance customer satisfaction and improve product rankings.

## Analysis

To calculate Value-For-Money, we developed a metric that incorporates product ratings, review count, and price. The formula applied was: The VFM calculation method aimed to balance three critical aspects: customer satisfaction, product popularity, and affordability. Customer satisfaction was determined using the average rating of each product, with higher values indicating positive reception by consumers. Product popularity was measured using the number of reviews, as a larger volume of feedback generally signifies higher consumer trust and engagement. However, to mitigate the disproportionate influence of highly popular products, a logarithmic transformation was applied to the review count, ensuring that excessively reviewed items did not overshadow others in the rankings. Finally, affordability was incorporated into the formula by including price as the denominator, meaning that products with lower prices and strong ratings naturally achieved higher VFM scores. This ensured that products offering a balance of quality and cost-effectiveness were prioritized in the analysis.Once the VFM scores were computed for all filtered products, the dataset was prepared for ranking and further interpretation. The results provided a standardized measure of value-for-money across different products, allowing for an objective comparison of consumer-perceived worth. After obtaining the VFM scores, the dataset was sorted in descending order to highlight the highest-value products. The ranking process allowed us to identify which products consistently delivered strong consumer satisfaction at competitive prices.

To calculate what percentage of high-rated products are actually good value for money, we established a benchmark, and we determined the 50th percentile (median) of VFM scores across all products. We then filtered out high-rated products (those with an average rating of 4.5 or higher) and identified which of these also had VFM scores above the median threshold, categorizing them as "good value-for-money" items. Finally, we computed the percentage of high-rated products that met this criterion. This method allows us to quantify how often highly rated products truly provide good value, revealing potential gaps between customer ratings and actual affordability.

To analyze which aspects in high-rated reviews influence customers' perceptions of value-for-money, we applied a structured natural language processing (NLP) pipeline using

PySpark. First, we loaded and preprocessed Amazon review data for both Electronics and CD & Vinyl categories, filtering reviews with a rating of 4 or higher to focus on positive customer feedback. We then cleaned the text by converting it to lowercase, removing punctuation, and tokenizing it into individual words. To ensure only meaningful words were considered, we removed common stopwords (e.g., “the,” “is,” “and”) using a filtering function.

Once the data was cleaned, we extracted and counted key product aspects mentioned in reviews, such as price, quality, features, design, value, performance, functionality, and durability. By analyzing word frequencies, we identified which aspects were most commonly discussed in high-rated reviews. Additionally, we examined sentiment by linking specific aspect mentions to high (4-5) or lower (below 4) sentiment scores, revealing which factors had the greatest impact on perceived value-for-money.

To further refine our analysis, we focused on top value-for-money products by computing a Weighted Value-for-Money (WVFM) Ratio using the formula:

WVFM Ratio = (log(Number of Related Products + 1)) / Price

We ranked products based on this metric and selected the top 50% percentile as high-value items. Finally, we matched reviews to these products, extracted relevant aspect mentions, and analyzed which attributes most influenced customers' perception of value. This methodology provides a data-driven approach to identifying key product features that impact value-for-money across different categories.

# Results and Discussion

The results of the VFM analysis for the Electronics category reveal several key trends. The top-ranking products based on VFM scores were primarily accessories such as protective cases, cables, and adapters. The highest-ranked product was the Poetic ThinShell Snap On Slim-Fit PC Hard Case for iPad (3rd Gen), which achieved the highest VFM score due to its combination of a strong average rating of 4.46, a substantial 520 reviews, and a competitive price point. This trend was also observed in other top-ranking products, including Canon PSC-55 Deluxe Leather Compact Case and Garmin 4.3-Inch Carrying Case, both of which had high review counts and strong customer ratings relative to their prices.

| title | average_rating | number_of_reviews | value_for_money_ratio |
|---|---|---|---|
| Poetic ThinShell Snap On Slim-Fit PC Hard Case for The New iPad(3rd Gen) / iPad 3 /iPad HD - Black (3 Year Manufacturer Warranty From Poetic) | 4.457692307692308 | 520 | 2788.620883996982 |
| Canon PSC-55 Deluxe Leather Compact Case for SD430, SD500, SD550, SD600, SD630, SD700IS, SD800IS, SD850 IS, SD900,SD950IS & SD870IS 300 HS , 100 HS Canon Digital Cameras models | 4.338738738738739 | 555 | 2742.416225685658 |
| Garmin 4.3-Inch Carrying Case | 4.4896445131137558 | 647 | 2647.5911564989365 |
| Poetic(TM) Slimbook Leather Case for Coby Kyros MID7012 7-Inch Android Tablet (3 Year Manufacturer Warranty From Poetic) | 4.372384937238493 | 239 | 2396.3463074863516 |
| Poetic (TM) PU Folio Case for Latest Generation Amazon Kindle 4 Wi-Fi 6" E Ink Display (4th Generation 6" Kindle Wi-Fi w/o Keyboard, NON-TOUCH Version) Dark Brown | 4.28793774319066 1 | 257 | 2381.074499059775 |
| Premium High Resolution 10FT / 3m 24K GOLD HDMI TO DVI M/M CABLE FOR HDTV PLASMA DVD | 4.376470588235295 | 170 | 2250.2339329635174 |
| Metra 40-CR10 Chrysler 2002 Antenna Adapter Cable | 4.57 | 100 | 2109.110076196456 |
| Neewer Retractable 3.5mm Aux Auxiliary Cord for 3.5 mm headphone jack Devices-Black | 3.7859922178988326 | 257 | 2102.3461774819975 |
| HDE® Mini 3.5mm Flexible Microphone for PC | 3.5195822454308809 | 383 | 2094.3775876992836 |
| Belkin 6ft / 3 Prong Notebook Power Cord | 4.640449438202247 | 89 | 2088.1139256701117 |

only showing top 10 rows

Figure 1. Top 10 best value for money products in the “Electronics” category

After assigning the value-for-money scores, we wanted to see out of the top-rated products, i.e., products having a rating greater than 4.5, how many were actually not only top-rated but also provided the best value for money. We set a modest threshold of choosing products whose VFM score (value-for-money score) fell in the top 50 percentile. Surprisingly, only 40% (approximate) of products were actually value for money. This means high ratings don't always mean a product is worth the price. Many popular products might be overpriced, and buyers could be influenced by branding, marketing, or reputation rather than the actual price-to-performance ratio. This creates a gap between what people think is valuable and what actually is, as 60% of highly-rated products may not be worth their cost despite positive reviews. This also suggests that many shoppers rely too much on ratings, where factors like brand loyalty and appearance may outweigh actual product quality. To make smarter choices, consumers should also look at value-for-money scores, durability, and alternatives, rather than just high ratings. This creates an opportunity for brands to offer better-priced quality products, and e-commerce platforms could use a ranking system based on VFM scores to help shoppers make better decisions. A new recommendation model combining customer satisfaction with VFM ratings could help buyers get the best value without falling for misleading ratings.

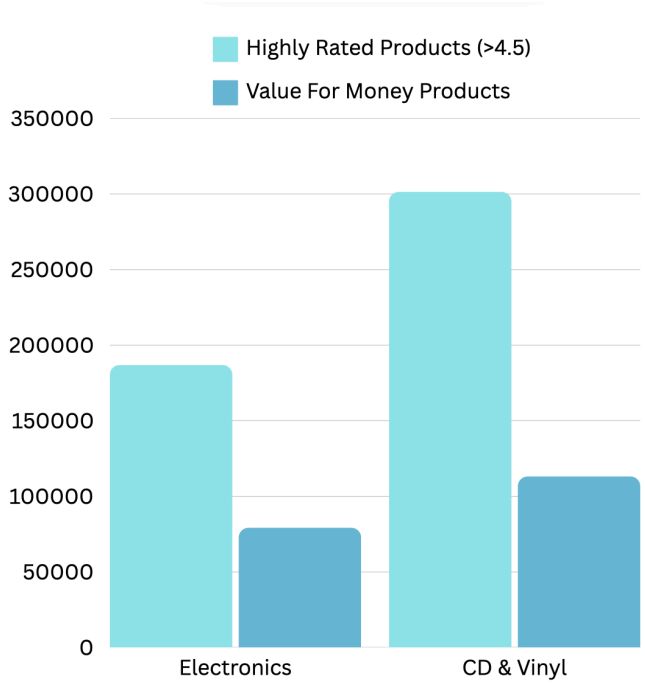


Figure 2. Comparison of Highly Rated and VFM Products in Electronics and CD & Vinyls

To understand what aspects mentioned in the high-rated reviews of the "Electronics" category most influence customers' perceptions of value-for-money Customer discussions around products primarily revolve around price, which stands out as the most frequently mentioned factor, with 1,136,383 references. This indicates that cost and its alignment with perceived value play a significant role in purchasing decisions. Following closely, quality emerges as another critical consideration, with 1,032,424 mentions, suggesting that customers consistently evaluate

the build and performance of products when forming opinions. Specific features also hold considerable weight, with 206,341 mentions, indicating that users often highlight particular functionalities that contribute to their satisfaction or dissatisfaction. Similarly, .design is an essential aspect, with 170,720 discussions focusing on aesthetics and structural appeal, demonstrating that visual and ergonomic factors influence customer perceptions. The concept of value, directly tied to "value-for-money," appears in 143,829 mentions, reinforcing its importance in consumer decision-making. Additionally, performance is a key metric, mentioned 141,076 times, emphasizing the significance of efficiency and effectiveness in product reviews. While functionality has fewer mentions at 41,399, it remains an important aspect that users focus on when assessing a product's usability. Lastly, durability, though mentioned only 22,883 times, is still a vital factor, especially for products expected to have a long lifespan. Together, these factors paint a comprehensive picture of what matters most to customers when evaluating a product.While in the case of CD and Vinyl ,The sentiment distribution across different product attributes highlights how customers perceive key aspects of their purchases. Design generally receives positive feedback, with 2,327 positive mentions compared to 547 negative ones, indicating that most customers appreciate the product's aesthetics and structure. Durability, while mentioned less frequently, has a positive sentiment (65 mentions) significantly outweighing negative feedback (11 mentions), suggesting that when durability is discussed, it is mostly in a favorable context. Features stand out with 50,388 positive mentions versus 6,571 negative ones, demonstrating that customers tend to appreciate the functionality and offerings of a product. Functionality, though discussed less often, also skews positively (24 positive vs. 12 negative mentions), reinforcing the idea that usability and operational aspects are generally well-received. A critical factor, performance, has a strong positive sentiment (73,149 mentions) compared to 12,255 negative mentions, indicating that customers value products that meet or exceed their expectations in terms of efficiency and effectiveness. Price, a frequent topic, shows 37,703 positive mentions versus 5,895 negative mentions, suggesting that while affordability and value-for-money are praised, a significant number of customers also express concerns over cost. Quality, another major factor, exhibits a strong positive skew, with 108,281 mentions appreciating it, compared to 26,571 complaints, reaffirming that product excellence remains a primary driver of customer satisfaction. Lastly, value, directly linked to affordability and quality perception, has 9,329 positive mentions against 2,509 negative ones, showing that while many customers feel they are getting a good deal, some still perceive the pricing as misaligned with the product's worth.
Overall, while most product aspects receive overwhelmingly positive feedback, features, performance, and quality stand out as the most appreciated attributes, whereas price and quality still generate a notable share of dissatisfaction.

When comparing the value-for-money of Electronics and CD & Vinyl products, a clear difference emerges. Electronics tend to offer better overall value, with an average value-for-money score of 2360.02, significantly higher than CD & Vinyl's 1185.01. The data also shows that Electronics have more consistent value, as reflected in the lower standard deviation (278.95 vs. 311.65). The boxplot reinforces this trend, showing that Electronics not only have a higher median value but also a more compact range, meaning customers generally agree on their worth. On the other hand, CD & Vinyl products vary much more, with some albums offering great value while

others fall short. This could be due to factors like niche demand, pricing, or how people perceive the worth of music versus tech gadgets. Overall, value-for-money in Electronics remains fairly stable, while in the CD & Vinyl category, it fluctuates considerably, making some products feel like a bargain and others not as much.

| title | average_rating | number_of_reviews | value_for_money_ratio |
|---|---|---|---|
| Gloria | 4.6415094339622645 | 53 | 1851.490708405305 |
| Fairweather Johnson | 4.379310344827586 | 29 | 1489.4898878313577 |
| Golden | 4.125 | 32 | 1442.3093691049232 |
| null | 5.0 | 10 | 1198.9476363991853 |
| Modulate | 2.9361702127659575 | 47 | 1136.650509585721 |
| Abacus Moon | 4.857142857142857 | 7 | 1010.0144631016343 |
| Freedom | 4.1 | 10 | 983.1370618473318 |
| Ken Holloway | 4.714285714285714 | 7 | 980.3081553633511 |
| Don't Let This Moment End / Oye / Disco Medle | 4.428571428571429 | 7 | 920.8955398867845 |
| Drew's Famous Steel Drums of the Island | 5.0 | 5 | 895.8797346140274 |

only showing top 10 rows

Figure 3. Top 10 best value for money products in the “CD & Vinyl” category

| Category | Mean | Std Dev | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|---|
| CD & Vinyl | 1185.01 | 311.65 | 895.88 | 981.02 | 1073.33 | 1366.72 | 1851.49 |
| Electronics | 2360.02 | 278.95 | 2088.11 | 2104.03 | 2315.65 | 2584.78 | 2788.63 |

Table 3. Comparative analysis of the “Electronics” and the “CD & Vinyl” category

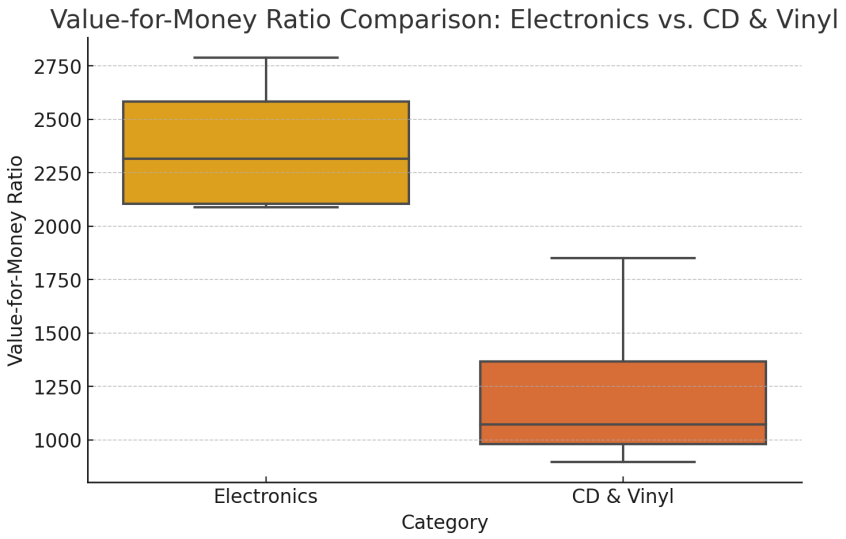


Figure 4. Value for money ration comparison of the “Electronics” and the “CD & Vinyl” category