Technical Report: Final Project DS 5110: Introduction to Data Management and Processing

Team Members:
Runying Chen & Bohan Cao
Khoury College of Computer Sciences
Data Science Program
chen.runy@northeastern.edu & cao.boh@northeastern.edu

November 17, 2024

Contents

1 Introduction

The rapid growth of e-commerce has changed how consumers discover and purchase products. Among various online platforms, Amazon stands as a global leader which consistently driving the majority of online retail sales. Its vast product range and dynamic nature of its bestseller lists provide valuable insights into what customers are buying, what products are performing well, and current market trends. Understanding these trends is essential not only for businesses looking to optimize their strategies but also for consumers aiming to make informed purchasing decisions.

In this project, we collected data from Amazon's bestseller lists, processed it for consistency, and conducted in-depth analyses using visualizations and statistical techniques. Our most recent data, gathered on October 16, 2024, provides a comprehensive snapshot of Amazon's bestseller landscape. By analyzing top-performing products across various categories, we aim to uncover patterns related to consumer preferences, pricing tactics, and product attributes that drive high sales. This study aims to serve as a reference point for businesses, analysts, and consumers interested in navigating the competitive world of e-commerce.

2 Literature Review

The study of e-commerce, particularly on platforms like Amazon, has attracted significant academic attention due to the platform's substantial influence on global retail dynamics. According to the research article "E-commerce Consumers Consumers Interest in Product Features Presented in Online Offerings", the authors indicate that "Gaining and responding to consumer interest in offered products is critical in e-commerce communication" (Trzebiński & Marciniak, 2023, p. 224). However, Amazon does not openly provide access to all of its databases, which presents challenges for researchers. Much of the available data about Amazon online is often outdated or overly simplistic, limiting its usefulness for in-depth analysis. This lack of access to comprehensive data creates difficulties in achieving the detailed insights needed for projects like ours, where analyzing current trends is crucial.

Because Amazon's bestseller rankings change frequently, relying on static datasets may not adequately capture the fast-paced changes in consumer preferences and market dynamics. As Zhang (2024) notes, "Recommendation systems generated by big data technology can personalize recommendations to consumers and improve the e-shopping experience for both consumers and suppliers" (p. 2). Therefore, decisions made based on static datasets can lead to outdated conclusions that do not accurately reflect present-day consumer behavior. Furthermore, existing studies have often concentrated on a narrow selection of categories or products, which does not provide a broad understanding of trends across Amazon's diverse marketplace. Another gap in existing research is the limited use of real-time analytics, which could offer businesses actionable insights that are relevant at the moment.

As the e-commerce landscape continues to evolve, there is an increasing need for more dynamic and interactive methods that allow stakeholders to quickly adjust to market shifts. To address these gaps, our project emphasizes real-time data collection, thorough data preprocessing, and the use of interactive visualizations to gain deeper insights into Amazon's bestseller lists. By leveraging the ScrapeHero API, we are able to gather the latest data across 36 categories, ensuring that our analysis remains up-to-date. Our approach goes beyond conventional analysis by using dynamic tools like scatter plots, word clouds, and data tables. These visualizations enable users to explore consumer behavior, pricing strategies, and product performance in a more interactive and comprehensive way, providing richer insights into the competitive landscape on Amazon.

3 Methodology

3.1 Data Collection

To ensure the accuracy of our analysis in reflecting the latest trends on Amazon, we utilized the ScrapeHero API to collect real-time data. This reliable online tool allowed us to extract structured data from Amazon's bestseller listings across 36 categories, including Electronics, Beauty, Books, and Home & Kitchen. To efficiently gather product details, we developed Python functions to automate the extraction process. We defined functions to pull unique ASINs from a CSV file, to make batch API requests to fetching product details, and to transform JSON format data into a structured csv format. Our automated pipeline efficiently handled large volumes of data, enabling us to refresh our dataset regularly. Through this API, we obtained detailed information on top-selling products such as ASIN (Amazon Standard Identification Number), product titles, ratings, prices, review counts, and product descriptions. Our real-time approach enabled us to gather the most current data as of October 16th, 2024.

3.2 Data Preprocessing

The initial dataset obtained from the ScrapeHero API contained noise, missing values, and inconsistencies that required addressing before meaningful analysis could be conducted. The team loaded the dataset into a Pandas DataFrame to explore its structure, which consisted of 25 columns with various data types. To enhance data quality, missing values were manually addressed by reviewing product URLs and using Python functions to fill in gaps for key variables such as `sale_price`, `ratings_count`, `rating`, `is_prime`, and `is_aplus_page`. Redundant or non-useful columns were identified and removed to streamline the dataset. Data type adjustments were made, including converting certain columns to Boolean types and transforming string-based rating columns into floats. Additionally, dollar signs in the `sale_price` column were removed and values converted to floats. Missing entries in specific columns were filled with placeholders, while continuous variables such as `ratings_count` and `sale_price` were scaled using z-score and min-max normalization techniques. This step helped standardize the data, making it easier to identify outliers and perform further statistical analysis.

To further enrich the dataset, to estimate the potential revenue generated by each product, a new column, `estimated_revenue`, was added. This was calculated by multiplying `ratings_count` by `sale_price`. This metric provided insights into the sales performance of products, allowing us to rank products based on their estimated financial impact. In addition, we included a column to assess the price-to-quality ratio of each product. A lower ratio indicates better value in relation to quality, while a higher ratio suggests that a product may be overpriced relative to its quality. This approach enabled us to conduct a more comprehensive evaluation of product value and gain deeper insights into consumer preferences and market dynamics.

3.3 Analysis Techniques

3.3.1 Correlation Analysis

We began our analysis by generating a correlation matrix to explore the relationships between various numeric features, such as ratings count, sale price, and estimated revenue. This was visualized using a heatmap to identify significant correlations that could indicate multicollinearity. For instance, high correlation between z-score features (like ratings count and sale price) suggested redundancy, which could impact regression models. As a result, we decided to focus on non-regression-based analyses to avoid unreliable coefficient estimates.

3.3.2 Descriptive Analytics

Category-level insights: To gain deeper insights into market segments, we analyzed the top-rated categories based on the total sum of ratings and average rating. Dual-axis charts were employed to compare the total ratings count against the average rating for each category, providing a balanced view of popularity and customer satisfaction. We also looked at categories with the highest estimated revenue, which was visualized using bar charts to highlight which product segments generate the most income on Amazon.

Product-level insights: We identified the top-rated products by sorting the data based on ratings count and average rating. We visualized the top products using horizontal bar charts to highlight those with the highest customer engagement. On the other hand, we analyzed products with the highest estimated revenue. By calculating the estimated revenue as the product of ratings count and sale price, we identified the top-selling items and visualized them to highlight which products drive significant sales volume on Amazon. The top products with the best price-to-quality ratio were identified and visualized to help consumers find high-value purchases. To compare product popularity with pricing, we created scatter plots using z-scores of sale price and ratings count. This helped in visualizing how pricing strategies impact customer ratings and purchase decisions.

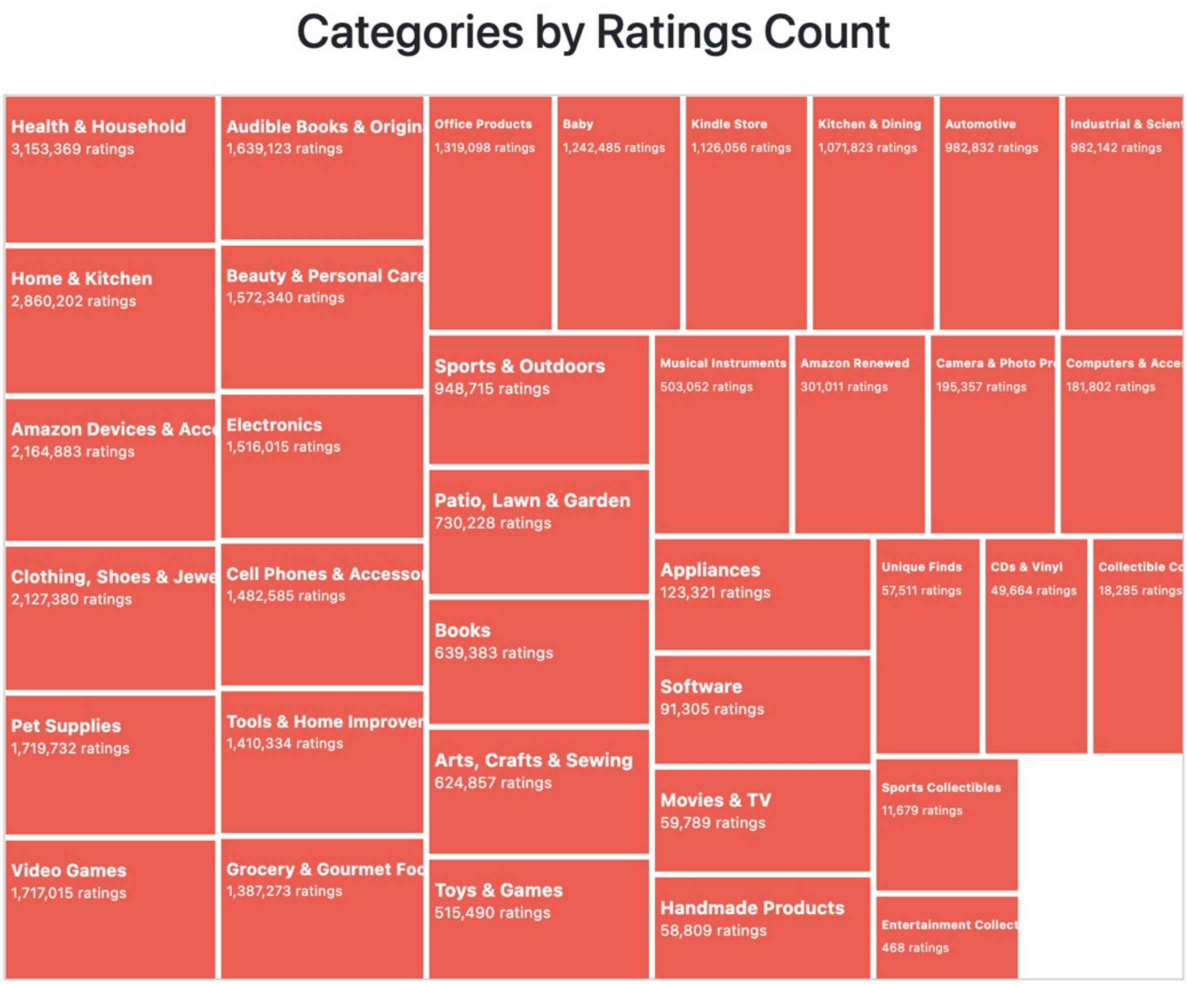
3.3.3 Textual Analysis

We utilized the Natural Language Toolkit (NLTK) to perform an in-depth textual analysis of Amazon bestseller product titles. The main objective was to gain insights to help businesses optimize their product titles to better align with consumer search behaviors and preferences, ultimately driving higher engagement and sales. To achieve this, we conducted text tokenization by combining product titles into a single string, breaking them into individual words, and using NLTK's stopwords filter to remove irrelevant terms. The WordNet Lemmatizer was employed to standardize words to their base forms, ensuring consistency. Additionally, we cleaned the text by removing punctuation, filtering out non-alphabetical characters, and converting all words to lowercase. We focused on word frequency to identify commonly used terms in product titles. This revealed popular keywords such as "pack," "Amazon," "iPhone," and "water," highlighting key product features that attract consumer attention. We further expanded our analysis by generating bigrams and trigrams to uncover meaningful word combinations. By analyzing product names, we aimed to understand what drives consumer interest and how products are positioned in the highly competitive Amazon marketplace.

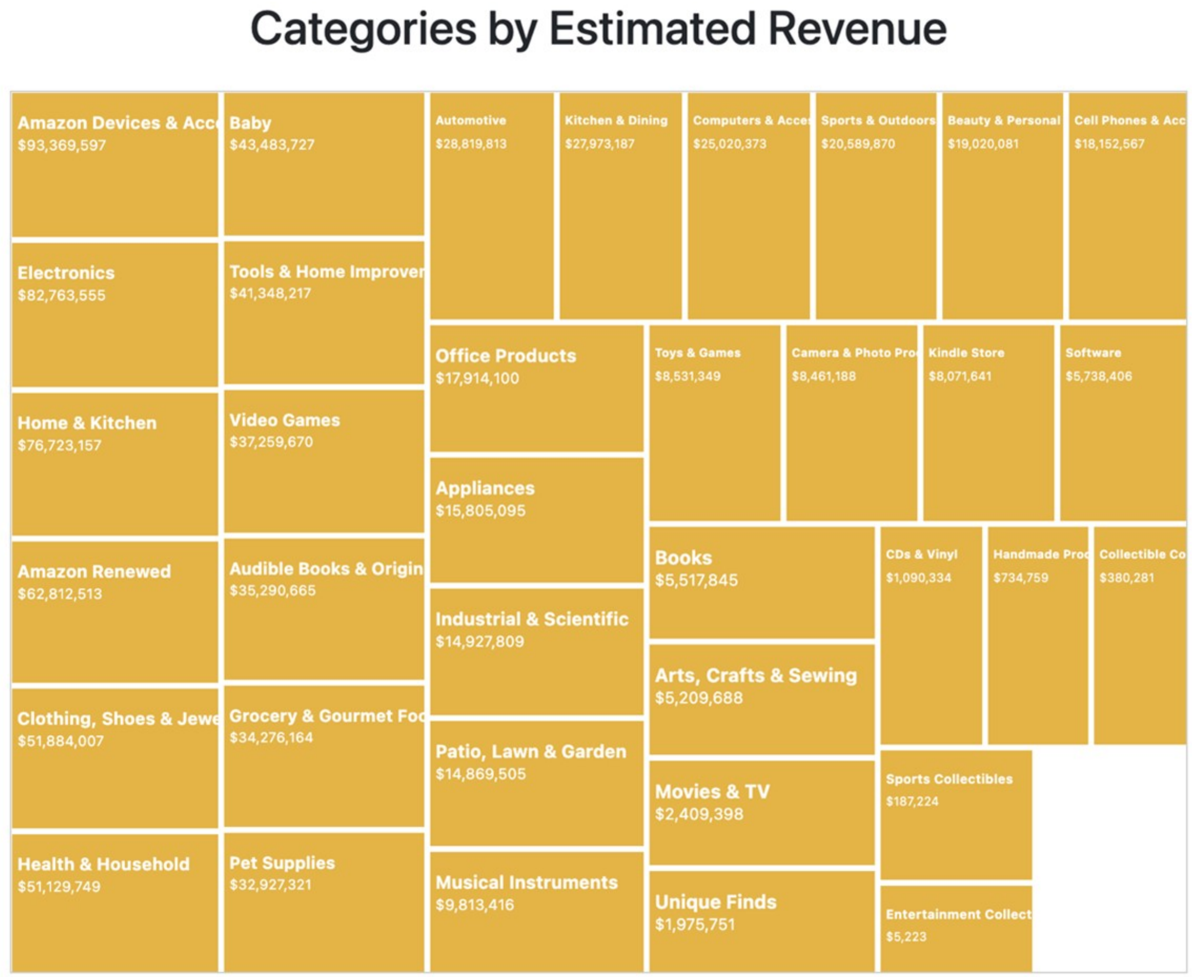
3.3.4 Visualization Techniques and Tools Used

To effectively communicate trends, patterns, and actionable insights, we utilized various JavaScript libraries, including D3.js, Recharts, and React components, each serving a specific purpose in enhancing user interaction and data comprehension. The project was styled with a blend of CSS and Bootstrap for a sleek, responsive design, and it was deployed using Netlify for easy remote access and scalability. We implemented filterable bar charts using D3.js and Recharts to allow users to analyze product performance metrics such as ratings count and estimated revenue across different categories. The charts were designed to be interactive, enabling users to explore data dynamically, filter categories, and focus on specific product segments. Treemaps were created to visualize hierarchical data, allowing users to understand the distribution of products across various categories. The treemap's color-coding and size indicators helped users quickly grasp which categories had the most products and which performed better in terms of sales and ratings. We utilized scatter plots to highlight the relationship between product prices and ratings, providing insights into consumer preferences and pricing strategies. By plotting products as dots, we allowed users to identify outliers, high-value products, and trends in pricing. Custom tooltips were implemented to show additional product details like sale price, ratings count, and product images when hovering over each dot, enhancing user engagement and interactivity. A Wordcloud was dynamically generated using text analysis to visualize frequently occurring terms in product titles. This helped reveal popular product features and attributes that attract consumer attention.

4 Results

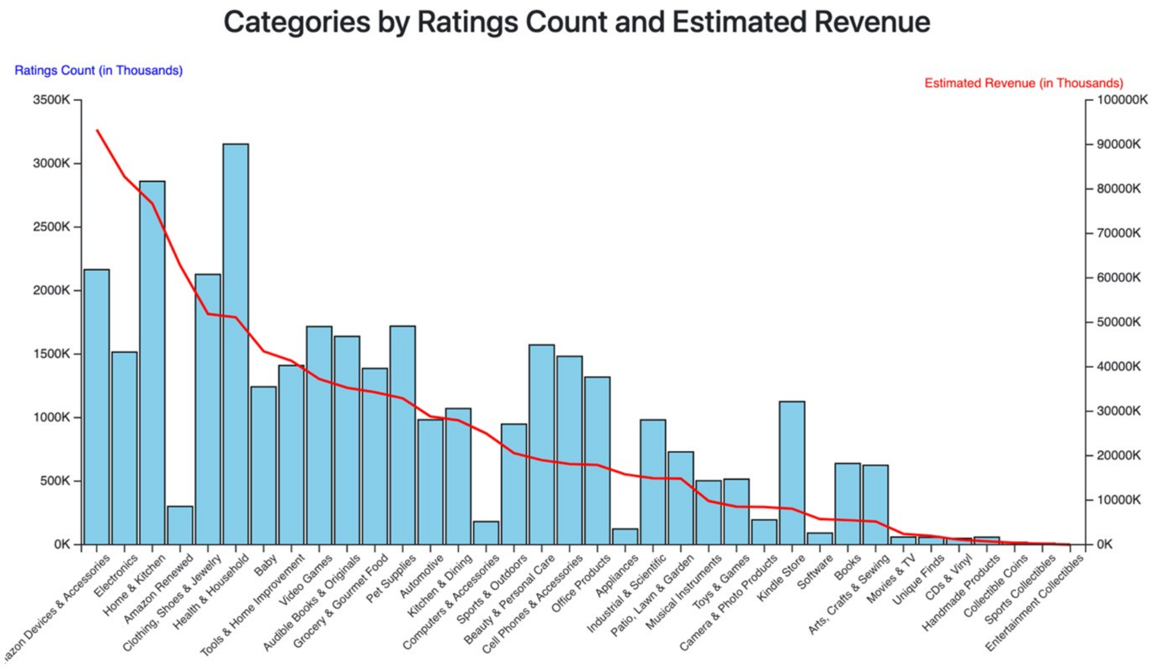


This treemap highlights the number of ratings across Amazon categories, which serves as a proxy for customer engagement and product popularity. Categories like Health & Household, Home & Kitchen, Amazon Devices & Accessories, Clothing Shoes & Jewelry have the highest number of ratings, suggesting these segments have the most customer interaction and the highest levels of consumer satisfaction.

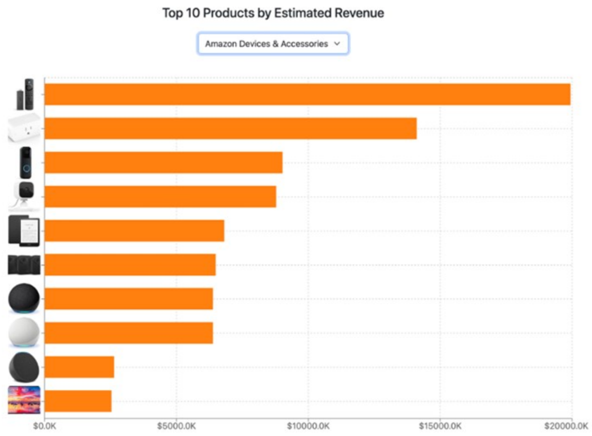


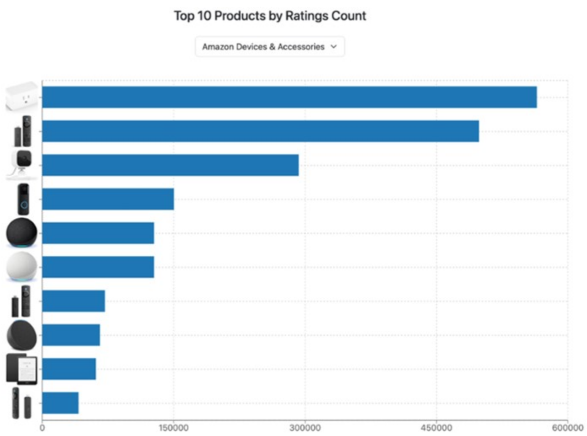
This second treemap visualizes the estimated revenue generated by various Amazon categories. Categories such as Amazon Devices & Accessories, Electronics, Home & Kitchen, Amazon Renewed, and Clothing Shoes & Jewelry dominate in terms of revenue, indicating their strong performance in sales.

When comparing the two treemaps, we noted that Amazon Devices & Accessories, Home & Kitchen, and Clothing Shoes & Jewelry are leading both in popularity and driving Amazon's sale. The Health & Household category, while not leading in revenue, shows the highest customer engagement based on ratings count. This indicates that, although individual items may have lower prices, the volume of sales and customer interest is substantial.

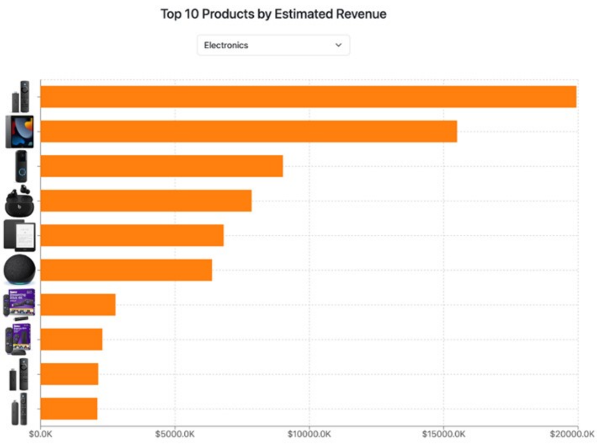


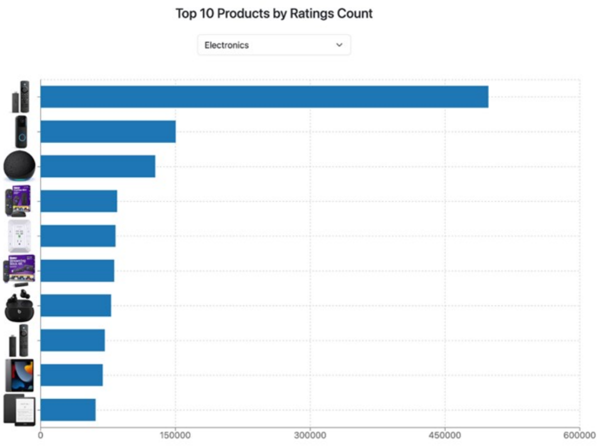
The dual-axis chart offers a detailed comparison of the ratings count (displayed on the left y-axis) and estimated revenue (on the right y-axis) across various Amazon product categories. The analysis reveals that categories like Amazon Devices & Accessories, Electronics, Home & Kitchen, Clothing, Shoes & Jewelry, and Health & Household dominate in terms of both popularity and sales. These categories not only attract a high number of customer ratings but also generate significant revenue, indicating a strong alignment between consumer interest and sales performance. This suggests that products within these categories are particularly effective at converting customer engagement into revenue. Nevertheless, we wanted to dive deeper to understand the key factors driving these results.


Top 10 Products by Estimated Revenue
Amazon Devices & Accessories
$0.0K
$5000.0K
$10000.0K
$15000.0K
$20000.0K

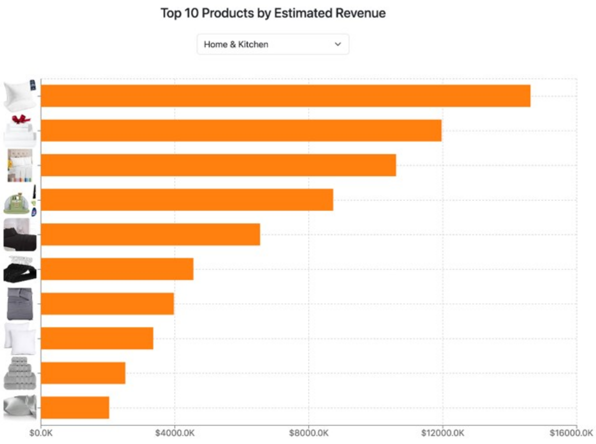

Top 10 Products by Ratings Count
Amazon Devices & Accessories
0
150000
300000
450000
600000

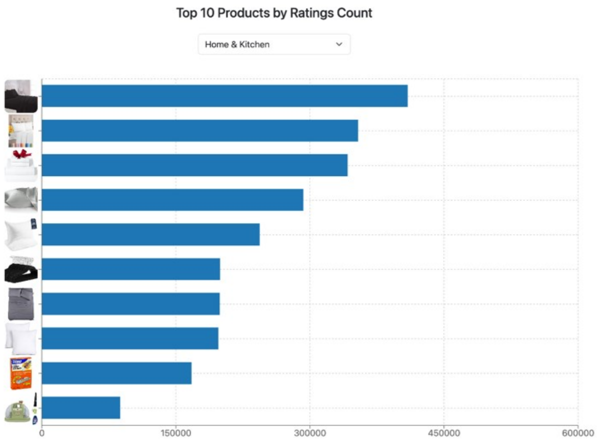
To uncover what contributes to the success of the Amazon Devices & Accessories category, we analyzed the top-performing products by estimated revenue and ratings count. As illustrated in the subsequent bar charts, popular items like TV remotes, Amazon Echo Dot, Smart Plugs, and the Amazon Kindle are among the primary drivers of revenue in this category. These products stand out not only for their technological innovation but also for their ability to meet consumer demands for convenience, smart home integration, and entertainment.



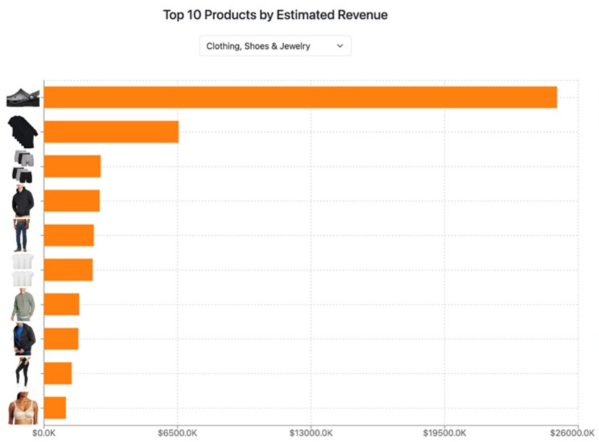


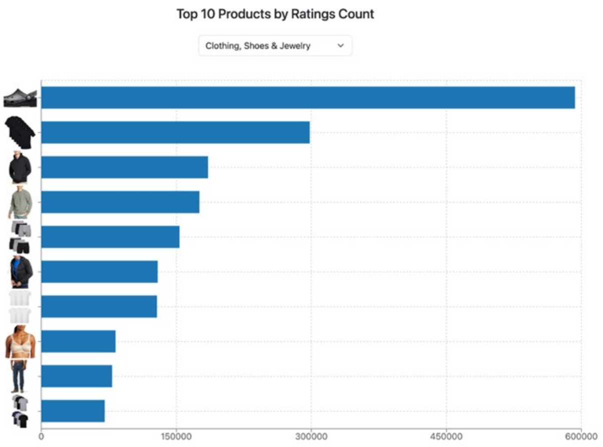
In the Electronics category, it is not surprising to see some overlap with the top-performing items in the Amazon Devices & Accessories segment. Products like the Amazon Echo Dot, Kindle, and Fire TV Stick continue to dominate both in terms of revenue and ratings, demonstrating their consistent appeal to consumers. The Fire TV Stick remains a standout performer, driving significant sales and customer engagement across both the Devices & Accessories and Electronics categories. Its versatility and ease of use make it a favorite among consumers looking for seamless streaming solutions.



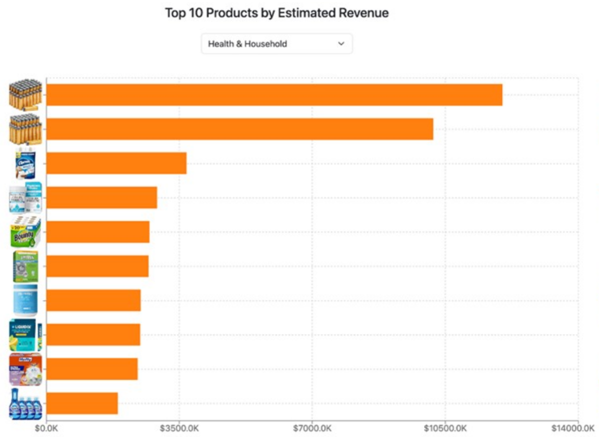


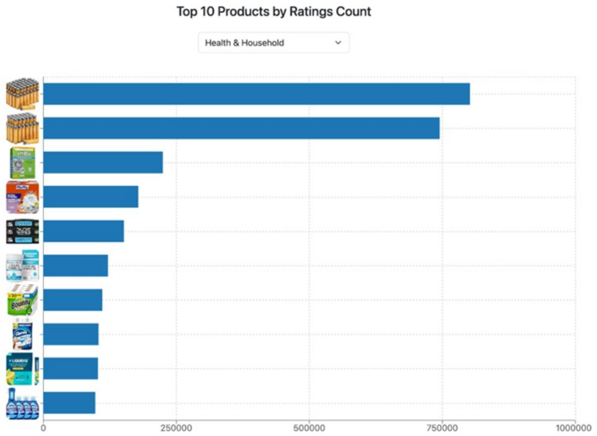
In the Home & Kitchen category, our analysis reveals a strong consumer preference for bedding products, such as bed sheets, pillows, and pillowcases. These items consistently rank high in both estimated revenue and ratings count, indicating their popularity among shoppers. This trend suggests that consumers are increasingly investing in products that enhance their sleep quality and overall comfort.



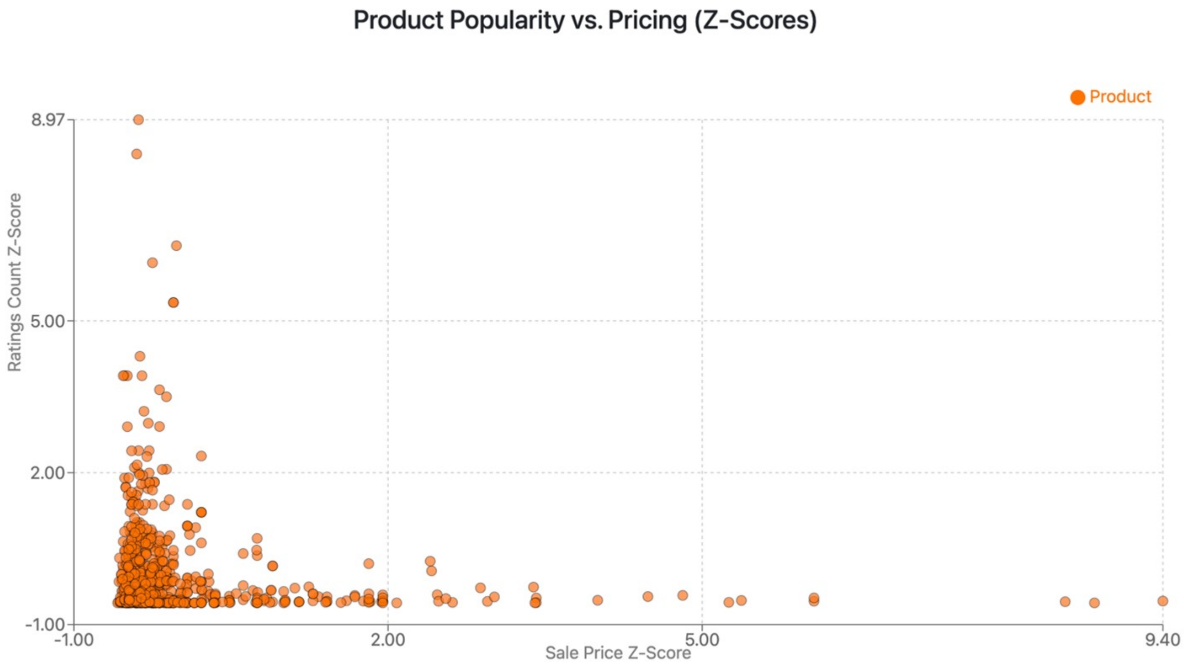


In the Clothing, Shoes & Jewelry category, the data reveals that clogs are the most popular product in terms of both sales and customer ratings. Interestingly, despite the general assumption that women drive most clothing purchases, our analysis shows that men's clothing outperforms in terms of sales on Amazon. The charts highlight a clear trend where men's fashion items, particularly jackets, hoodies, and casual wear, have higher revenue and rating counts compared to women's clothing. This suggests that male consumers on Amazon are increasingly purchasing fashion essentials, driven by convenience and the availability of diverse options online. The strong performance of men's clothing also points to a shift in consumer behavior where men are becoming more engaged in online fashion shopping, challenging traditional market perceptions.

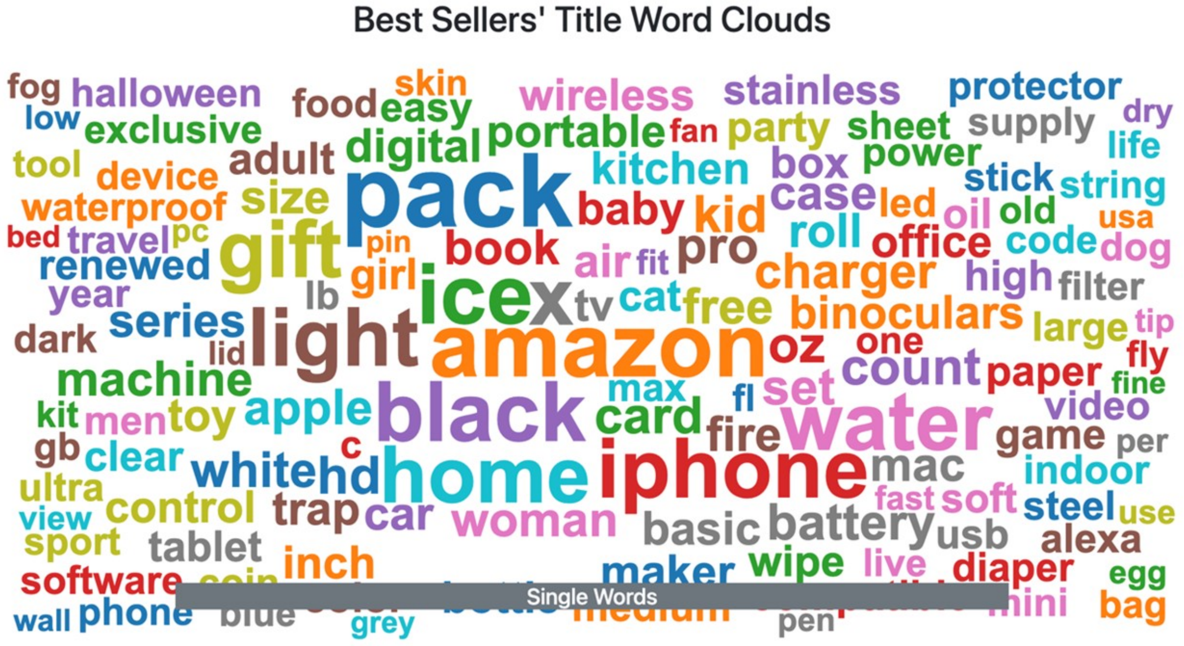




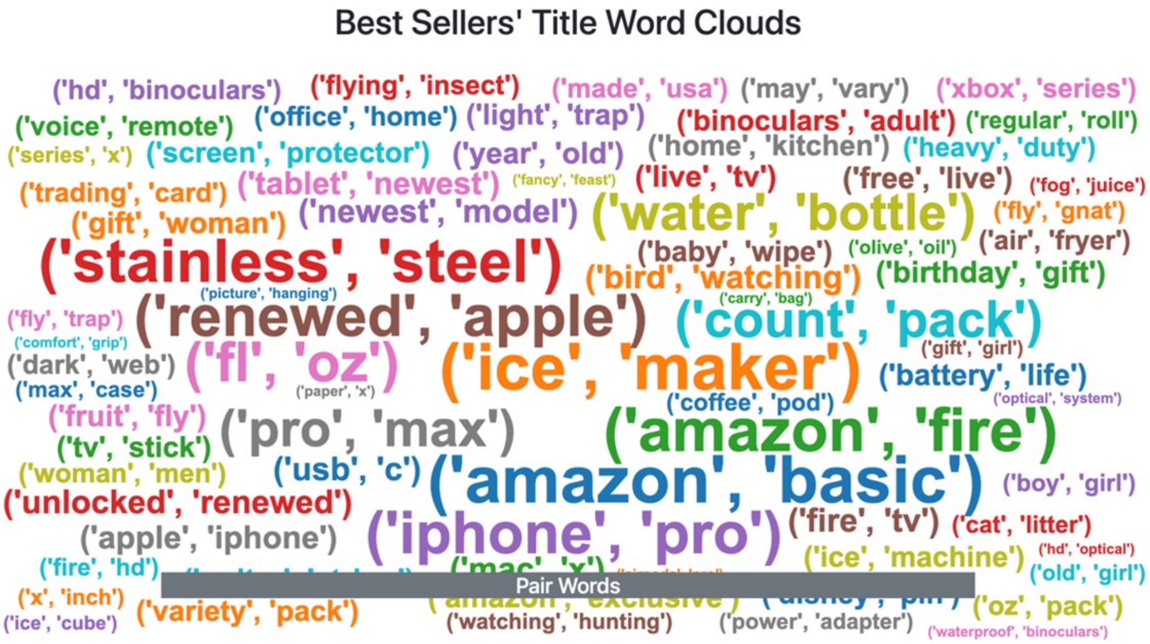
In the Health & Household category, the data clearly highlights that Amazon Basics High-Performance Batteries significantly dominate both revenue and customer ratings. The performance of these batteries is so outstanding that their total revenue exceeds the combined earnings of all other top products in this category. This trend indicates that consumers highly trust and prefer Amazon's in-house brand when it comes to household essentials like batteries, which are frequently needed and replaced.



In addition to the top categories, we created a scatter plot that illustrates the relationship between product prices (on the x-axis) and the number of customer ratings (on the y-axis), both normalized using Z-scores. The chart reveals a clear trend: products with lower sale prices tend to receive a higher volume of customer ratings, indicating greater popularity. As the price increases (towards the right side of the x-axis), the number of ratings sharply decreases, showing that higher-priced items generally attract fewer buyers. The concentration of data points in the bottom-left corner suggests that the most popular products on Amazon are those that are priced competitively. This pattern reflects consumer preferences for budget-friendly items, especially in the highly competitive online marketplace.



Additionally, we generate wordclouds to provide insights into the key terms frequently used in Amazon's best-selling product titles, which highlights consumer interests and popular product features. The first word cloud, which focuses on individual words, shows that terms like "pack," "gift," "Amazon," "iPhone," "water," "home," and "black" are among the most prominent. These words suggest that consumers are particularly interested in multipacks (indicated by "pack"), technology (e.g., "iPhone"), and home-related items (e.g., "home," "kitchen"). The appearance of words like "gift" indicates that many best-selling products are popular gift items, potentially driven by seasonal shopping trends.



The second word cloud focuses on word pairs (bigrams), revealing more specific product associations. For instance, phrases like "stainless steel," "ice maker," "iPhone pro," and "Amazon fire" highlight popular product features and categories. The pairing of words like "renewed apple" and "count pack" suggests a high demand for refurbished Apple products and bulk purchases. Phrases such as "water bottle" and "newest model" indicate a preference for hydration-related items and up-to-date technology. Additionally, combinations like "birthday gift" suggest that many products are purchased for special occasions, further emphasizing the gifting trend.

5 Discussion

The results from our analysis show clear patterns in consumer behavior on Amazon. We found that categories like Amazon Devices & Accessories, Home & Kitchen, and Clothing, Shoes & Jewelry lead both in ratings and revenue. This means these categories are very popular and generate a lot of sales. Products in these categories, like the Amazon Echo Dot, Kindle, and Fire TV Stick, stand out because they meet customer needs for technology and convenience.

Interestingly, the Health & Household category shows high ratings but lower revenue compared to others. This suggests that even though customers are highly engaged with these products, their lower prices may limit the revenue. This could be because people buy more basic or lower-cost items like batteries, which are essential but not very expensive. Our results match the literature review, which highlights the importance of customer engagement in determining sales success. However, we observed that some categories, such as Health & Household, do not follow the same revenue trends as others. This is a slight difference from what was expected in the literature, where high engagement often leads to high sales.

The comparison between ratings count and estimated revenue helps explain this difference. Categories like Amazon Devices & Accessories, which are both highly rated and generate high revenue, align with the findings in previous studies, which show that products with strong customer engagement tend to perform well in terms of sales. However, we also found that price plays a big role in how many ratings a product gets. Cheaper products tend to get more reviews, while expensive products, although popular, do not generate as many ratings. This shows that price sensitivity is an important factor for consumer behavior on Amazon, something the literature review did not emphasize as much.

Looking at the word clouds, we can see that certain terms like “pack,” “Amazon,” and “iPhone” appear frequently in best-selling product titles. This suggests that customers are drawn to multipacks and technology-related items. These findings are consistent with consumer trends mentioned in the literature review, where convenience and technology play a significant role in purchasing decisions. However, we also noticed that products labeled as “gift” appear often, which points to seasonal shopping trends that were not fully explored in the literature.

Overall, our findings support the idea that customer engagement and product pricing are crucial for sales performance on Amazon. The discrepancy in Health & Household products, where high engagement does not translate into high sales, could be due to pricing strategies that favor larger, higher-priced items. This highlights the need for businesses to consider both customer ratings and pricing strategies when planning their sales strategies.

6 Conclusion

In this project, we set out to analyze Amazon’s best-selling products across various categories to uncover key consumer preferences, market trends, and factors driving sales performance. Our comprehensive analysis, using data collected on October 16, 2024, provided valuable insights into how different categories perform in terms of popularity and revenue. Categories like Amazon Devices & Accessories, Electronics, Home & Kitchen, and Clothing, Shoes & Jewelry were identified as the most influential sectors, showing strong alignment between customer engagement and sales volume.

Our findings highlighted that consumer demand is significantly driven by products that offer convenience, utility, and value. For instance, items such as TV remotes, smart home devices, bedding products, and budget-friendly essentials like Amazon Basics batteries demonstrated high sales volumes. Additionally, men's fashion items outperformed women's clothing in sales, challenging traditional assumptions about online apparel purchases. The price-to-quality analysis and scatter plots further revealed that lower-priced items attract more customer ratings, confirming a consumer inclination towards value-driven purchases.

Despite the strengths of our approach, there were limitations. The reliance on a single API for data collection may have restricted the scope of our analysis, as some data points were not fully representative of Amazon's extensive product range. For future research, integrating more diverse data sources, including social media sentiment analysis and real-time price tracking, could provide a richer understanding of consumer preferences. Expanding the study to include seasonal trends and the impact of marketing campaigns could further enhance the insights gained from analyzing Amazon's dynamic e-commerce landscape.

7 References

- Trzebiński, W., & Marciniak, B. (2023). Meaning or importance? E-commerce consumers' interest in product features presented in online offerings: The role of self-relevance and information processing. *Journal of Internet Commerce, 22*(2), 224–243. `https://doi.org/10.1080/15332861.2022.2042116`

- Zhang, H. (2024). An accuracy study of personalized recommendation system for e-commerce based on big data analysis. *Applied Mathematics and Nonlinear Sciences, 9*(1). `https://doi.org/10.2478/amns-2024-1923`

A Appendix A: Code

Due to the large size of the code used in this project, we are providing a link to the full code on GitHub. You can access and review all the relevant scripts and functions at the following location:

`https://github.com/RennieCh/Amazon-Best-Seller/tree/main/src`