**Analysing Purchase Trends by Product Category Over Time on Amazon**

**E-commerce.**

1. **Introduction.**

* Online shopping has exploded in popularity and Amazon.com is at the forefront as the biggest online store in the world. In 2024, Amazon's revenue hit $638 billion, which is an 11% increase from 2023. (Amazon.com, 2025) This growth shows a bigger trend: the COVID-19 pandemic which has pushed more people to shop online, leading to a steady rise in global e-commerce sales. This research highlights how real-time analytics can tackle important issues such as market saturation, changing consumer preferences, and staying competitive.
  1. **) Research objective and Rationale of dataset/database chosen:**
* The research objective is to look at how buying habits on Amazon change over time by product category, using the 'Amazon Product Sales Dataset 2023' from Kaggle. This dataset includes hundreds of thousands of Amazon products from 142 different categories, featuring details like category, price, customer ratings, and the number of reviews among others. (Ding, 2024).
  1. **) Explanation of why the industry is particularly suited for data analytics and how the chosen dataset reflects its current challenges**:
* E-commerce data analytics works well because it involves fast and high-volume updates. This data shows patterns and issues such as pricing and quality in different categories of products on Amazon. Analysing it helps address these issues and understand past performance for strategic decisions. (Yong & Noor, 2025)

1. **Literature Review:**

* Recent research indicates that the online shopping industry, particularly Amazon, has quickly embraced data analytics following the increase in online shopping after COVID-19. (Amazon.com, 2025) Amazon leverages real-time data to customize user experiences, improve logistics, and predict demand. Additionally, analytics play a crucial role in identifying fraud and enhancing operational efficiency. These findings highlight larger trends in e-commerce, where making decisions based on data is vital for staying competitive, which makes Amazon's sales data very important for further study and to find out about buying trends and business issues. (Denk Rimakka & Aras, 2023)
  1. **Key trends, challenges, and opportunities identified by previous research:**
* Some important trends are Amazon using machine learning to personalize experiences, adjust prices dynamically, and create recommendation systems (Meduri, 2023). Experts note that product categories like auto parts, food, health, and clothing are growing, while book sales are dropping. There are challenges too, like managing huge amounts of messy data and issues with predictive models, including overfitting and low accuracy in simpler approaches (Tsagkias et al., 2020)
  1. **The relevance of data chosen based on the literature:**
* Because the data includes over 142 types of products, it permits precise analysis of product trends. Literature confirms this importance, pointing out that product price, customer reviews and available inventory stock all raise sales (Ding, 2024). The shape of the dataset matches with these listed fields. (Piyushkumar-pk007, 2023)
  1. **Some relevant academic or industry sources within last 5 years**.

1. (Ding, 2024).
2. (Tsagkias et al., 2020)
3. (Rimakka & Aras, 2023)
4. (Amazon.com, 2025)
5. (Meduri, 2023)
6. (Robinson, 2024)
7. (Piyushkumar-pk007, 2023)
   1. **Connection between the research and target of data analytics:**

This research aims to provide straightforward insights into e-commerce analytics. By using these ideas, we can analyse the way sales trends on Amazon change by category to better understand how people purchase. Since predictions are not the goal, the analysis centres around trend visualization and figuring out how categories perform, helping solve the interpretability and data restrictions mentioned in previous research. (Tsagkias et al., 2020) This method aligns with the literature's demand for practical and easy-to-grasp insights into shopping habits. (Robinson, 2024)

1. **Research Design and Sampling Methodology:**

* To fill in the gaps in earlier studies, this research uses descriptive analytics to look at sales trends by product category over time. It analyzes important factors like price, discounts, customer ratings, and review counts on a monthly and quarterly basis to show how consumer behaviour changes across different categories. (Lokeswari-Kalla, 2024) This method steers clear of complicated predictive modelling issues like overfitting and low accuracy, offering straightforward insights into buying habits that match the research goals. The selected approach focuses on being clear and helpful, responding to the need for practical analytics in e-commerce. (Amazon.com, 2025)
  1. **) Q) - Describe data, how it was collected, and why it is useful in my research. Outline the sampling method and size, explaining the criteria for selecting the data.**
* The dataset I used is the Amazon Product Sales Dataset from 2023, collected from the Amazon website and shared on Kaggle. (Parab, 2023) It has more than 300,000 product listings in 142 different categories. The sample uses 10–15 major categories based on the amount of product or profit to give a true and clear representation of the analysis. Important missing details (e.g., price, category) were removed and minor categories were consolidated under “Other” to avoid skewed results. (Chatterjee et al., 2021)
  1. **Q- Explain what elements or aspects you check for data validity, completeness, and representativeness.**
* The integrity of the data was upheld by verifying the logical consistency of price and rating values, as well as removing entries lacking essential information. Completeness was confirmed by including all significant product categories, while representativeness was preserved by prioritizing widely recognized, high-volume categories and combining fewer common ones to mitigate bias and facilitate dependable analysis. (Hassan, 2024)

1. **Data Review and Correction:**

* The numbers like price, rating, reviews, stock, and discounts are important because they affect how people decide to buy things and show how well a product is doing. Looking at these numbers helps us see sales trends, customer satisfaction, and what is going on with the inventory, which is important for getting a grasp on how e-commerce works on Amazon.(Klimek & Funta, 2021) The information was gathered through web scraping methods, which systematically retrieve comprehensive product details including pricing, ratings, reviews, inventory status, and discounts directly from Amazon's platform. This approach enables the collection of up-to-date sales data without the need for surveys or questionnaires. (Cibhiboopath, 2024)

**4.1) Describe the steps to ensure data accuracy and reliability during collection. Describe if the dataset/database contains all the data you need to answer the research problem. If not, how you can rectify the missing data.**

* The data was gathered by scraping Amazon's website and saving it in a CSV file. (Karkavelraja, 2022) There could be errors from how the HTML was read, such as formatting problems, missing information, or outliers. To make sure the data is accurate, I will clean it up by organizing numbers, combining similar categories, getting rid of duplicates, and removing records that lack important details like price or category. I'll also check for consistency using summary stats and cross-tabulations. Plus, I will make sure that the counts of products and price ranges make sense (for example, 'Books' usually cost less than 'Electronics'). The dataset doesn't have direct sales numbers, so I based the analysis on ratings and stock levels. This should be considered when looking at the results. These steps help make sure the dataset is still reliable even without the sales data. (Chatterjee et al., 2021)

1. **Basic Visualizations and summarization:**

* To find deeper insights beyond just simple summaries, I will utilize these advanced visualizations:
  1. **Heatmap of Average Ratings by Category and Price Range**
* The heatmap shows how happy customers are with different categories and price ranges, indicating that Books priced at a lower range (4.27) and Electronics (4.50) receive the highest ratings, indicating a strong perception of value. Conversely, mid-priced Books and Clothing receive the lowest ratings.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Average of Rating** | **Column Labels** |  |  |  |
| **Row Labels** | **High** | **Low** | **Medium** | **Grand Total** |
| Books | 3.88 | 4.266666667 | 2.325 | 3.458333333 |
| Clothing | 2.6 | 2.933333333 | 2.425 | 2.627272727 |
| Electronics | 3.766666667 | 4.5 | 3.325 | 3.6375 |
| Home & Garden | 3.42 | 3 | 3.725 | 3.438888889 |
| Toys | 4.033333333 | 3.28 | 3.266666667 | 3.481818182 |
| **Grand Total** | **3.515** | **3.382352941** | **3.126086957** | **3.328333333** |

* 1. **Bubble Chart of Category Sales Volume vs. Average Price with Review Count as Bubble Size**
* This bubble chart shows the relationship between sales volume and average price, where the size of the bubbles represents the number of reviews. Electronics and Home & Garden have a lot of sales and reviews, showing that many customers are really interested and they are popular.
  1. **Scatter Plot of Discount Percentage vs. Sales Rank Over Tim**e
* The scatter plot illustrates the connection between discounts and product rankings over time. It shows that bigger discounts tend to lead to improved sales ranks, particularly during peak times. This pattern implies that discounts play a significant role in boosting short-term sales.
  1. **Histogram of Stock Availability Across Categories**
* The histogram shows how stock is spread out among different product categories. The Home & Garden and Clothing sections have the most stock, which could mean there's a lot available or that things aren't selling fast. In contrast, Electronics have lower stock levels, likely because they're in high demand or there's not enough available.
  1. **Provide statistical summaries related to the research problem and provide interpretations.**
* The data shows that Electronics have the best average rating at 3.64, which means customers are really happy with them. On the other hand, Clothing has the lowest rating at 2.63, hinting that there might be some quality issues. Books get the most reviews, averaging 361.92, which shows that people are really interested in them, even though some studies say book sales are going down. The average price of products is ₹158.59, with Electronics being the priciest. Home & Garden and Toys have the most stock available. These findings highlight trends in different categories: Electronics are great for customer satisfaction, Books are popular among buyers, and Clothing might need some improvements. All this helps in understanding how consumers behave on Amazon.

1. **Conclusion and Recommendation for Data Analytics Process:**

* Electronics are the most liked by customers, getting an average score of 3.64.
* Books have the most reviews, averaging 361.92, even though people say book sales are going down.
* Clothing is not doing well in customer satisfaction, with the lowest score of - 2.63.
* Toys and Home & Garden have the most stock, showing that these categories are popular.
* Prices change depending on the category, with Electronics being the priciest. **These results show clear trends in what customers like and how products are doing on Amazon.**

1. **Provide actionable recommendations for improving data collection method in future studies.**

* To improve how we gather data in the future, we should:

1. Add time-stamped information to better see trends over months or quarters.
2. Gather demographic and regional details to help with analysis.
3. Keep a close eye on actual sales numbers, not just things like reviews or how much stock we have.
4. Routine Quality Assessments: Perform regular data evaluations during the research process. (Hassan, 2024)
5. **Explain how the recommendations are based on data analytics findings, and how they may improve the gap identified in literature review.**

* These suggestions come from the challenges faced in the current study, like not having enough time and sales data, which made it tough to back up claims found in other research. By fixing these issues, future studies can connect their findings more closely with overall market trends, make the data more representative, and strengthen the trustworthiness of conclusions made from online shopping analytics. (Yong & Noor, 2025). This method of analysis can help with managing inventory, setting prices, and deciding where to place products on online shopping sites.

**References.**

Amazon.com, Inc. (2025). *2024 Annual Report*. <https://s2.q4cdn.com/299287126/files/doc_financials/2025/ar/Amazon-2024-Annual-Report.pdf>

Chatterjee, I., Zhou, M., Abusorrah, A., Sedraoui, K., & Alabdulwahab, A. (2021). Statistics-based outlier detection and correction method for amazon customer reviews. *Entropy*, *23*(12), 1645.

Cibhiboopathy. (2023). *Amazon-Sales-Dataset-EDA* [GitHub repository]. GitHub. <https://github.com/CIBHIBOOPATHY/Amazon-Sales-Dataset-EDA>

Denk Rimakka, A. I., & Aras, R. A. (2023). User segmentation based on purchasing habits and preferences on the Amazon platform using K-means clustering. *INSPIR: International Journal of Social Sciences and Humanities*, *13*(2), 103–109. <https://doi.org/10.35585/inspir.v13i2.63>

Ding, W. (2024). *Analysis of customer purchasing behavior in E-commerce model*. In Proceedings of CONF-MPCS 2024 Workshop: Quantum Machine Learning: Bridging Quantum Physics and Computational Simulations. <https://doi.org/10.54254/2753-8818/52/2024CH0163>

Hassan, M. (2024, March 26). Data verification – process, types and examples. *Research Method*. <https://researchmethod.net/data-verification/>

Karkavelraja, J. (2022). *Amazon sales dataset* [Data set]. Kaggle. <https://www.kaggle.com/datasets/karkavelrajaj/amazon-sales-dataset>

Klimek, L., & Funta, R. (2021). Data and e-commerce: An economic relationship. *Danube: Law, Economics and Social Issues Review*, *12*(1), 33–44. https://doi.org/10.2478/danb-2021-0003  
Licensed under CC BY-NC-ND 3.0

Kumar, P. (n.d.). *Amazon sales analysis* [Power BI dashboard]. GitHub. <https://github.com/piyushkumar-pk007/amazon-sales-analysis>

Lokeswari-Kalla. (2024). *Amazon-Sales-Dataset* [GitHub repository]. GitHub. <https://github.com/Lokeswari-Kalla/Amazon-Sales-Dataset>

Meduri, E. D. K. (2023). Marketing analytics and its application in e-commerce industry: Amazon. *International Journal of Management Research & Business Strategy*, *12*(3), Article 70. <https://ijmrbs.net/index.php/ijmrbs/article/download/70/62/117>

Parab, L. (2023). *Amazon products sales dataset 2023* [Data set]. Kaggle. <https://www.kaggle.com/datasets/lokeshparab/amazon-products-dataset>

SageSeller. (2024, October 22). *How Amazon uses data analytics to drive its e-commerce success*. SageSeller. <https://sageseller.com/blog/how-amazon-uses-data-analytics-to-drive-its-e-commerce-success/>

Tsagkias, M., King, T. H., Kallumadi, S., Murdock, V., & de Rijke, M. (2021). Challenges and research opportunities in eCommerce search and recommendations. *ACM SIGIR Forum, 54*(1), Article 2, 1–23. <https://doi.org/10.1145/3451964.3451966>

Yong, P. L., & Noor, N. A. M. (2025). Navigating the research trends in online purchase behavior: A bibliometric study. *Multidisciplinary Science Journal*, *7*(11), 2025512-2025512.