AMAZON SALES DATA ANALYSIS

IE6600 – COMPUTATION AND VISUALIZATION FOR ANALYTICS

FINAL REPORT

GROUP NUMBER 2

Aman Malawade- 002762202

Brinda Raj L - 002796473

Hardika Shroff – 002743794

Naga Sumanth Reddy Bareddy – 002778629

Viraj Patil - 002770574

PART 1: INTRODUCTION AND RESEARCH QUESTIONS

INTRODUCTION

Amazon, the world's largest online retailer, offers an extensive range of products across multiple categories such as electronics, fashion, home goods, and more. In today's competitive business landscape, it is essential to gain insights into which products are performing well and why, to stay ahead of the competition and ensure customer satisfaction. This project aims to analyze Amazon's sales data to identify the top-performing categories and products, as well as the factors that influence their sales. Our analysis can be useful for Amazon and other businesses to optimize their product offerings, marketing strategies, and inventory management, leading to improved sales and customer satisfaction.

RESEARCH QUESTIONS

- Which product categories and subcategories are the most popular on Amazon?
- What factors influence a product's sales on Amazon? (e.g., pricing, customer ratings, discount effectiveness, product images, and links)
- Which products have the highest sales performance, and what can we learn from their success?
- How can Amazon sellers and marketers use the insights gained from this analysis to improve their sales and customer engagement?

This final project report serves as a comprehensive record of our Amazon Sales Data Analysis project, showcasing our ability to apply a range of data analytics skills in a cohesive manner. Our objective was to gain valuable insights into Amazon's product sales performance by identifying the most successful products, categories, and subcategories, as well as analyzing data to understand pricing strategies, sales performance, customer ratings, discount effectiveness, and the impact of product images and links. These insights have the potential to optimize marketing strategies, enhance customer engagement, and increase sales for Amazon sellers and marketers, thereby providing a competitive edge in the marketplace.

This report provides an overview of the entire analysis, beginning with research questions and data collection, followed by cleaning, analysis, and visualization. The results of the analysis answer the research questions and offer practical insights. The report will illustrate how these insights can be used to make informed decisions in the marketplace.

PART 2: SUMMARY OF RESULTS

- Collated 7 different datasets into one to obtain the main dataset: To conduct a comprehensive analysis of Amazon's sales data, we gathered multiple datasets from different sources and consolidated them into one primary dataset. This allowed us to have a more complete and accurate picture of the products and categories available on Amazon, as well as the sales performance of each.
- Conducted Exploratory Data Analysis on the cleaned dataset: To gain insights into the data and identify patterns and trends, we performed exploratory data analysis (EDA) on the

cleaned dataset. This involved visualizing the data through graphs, charts, and histograms to better understand the distribution of variables and to identify outliers.

- Identified and removed duplicates and handled missing values to improve data quality: Before conducting any analysis, we took the necessary steps to clean the dataset by identifying and removing any duplicates and handling any missing values. This helped to ensure that the analysis was based on accurate and complete data.
- Analyzed the influence of consumer interaction with the products to draw data-informed business insights: By analyzing consumer interactions with the products, such as ratings and reviews, we were able to draw insights on how customers engage with products on Amazon. This information can be used by Amazon and its sellers to optimize product listings and improve customer engagement.
- Identified key factors influencing sales performance on Amazon: Through statistical analysis, we identified the key factors that influence sales performance on Amazon. These factors included product ratings, price, discount percentage, and product images.
- Examined interrelationship between different variables: We examined how different variables were related to each other, such as the relationship between ratings and sales performance or the relationship between product price and discount percentage. This allowed us to identify trends and patterns in the data and draw insights on how different variables impact each other.
- Analyzed distribution of variables and identified patterns and trends: By analyzing the distribution of variables, we were able to identify patterns and trends in the data. For example, we found that certain product categories were more popular than others, and certain pricing was more effective in increasing sales.
- Identified outliers using statistical measures and visualizations: Through statistical measures and visualizations, we identified outliers in the data. These outliers can provide valuable insights, such as identifying high-performing products or categories that are underrepresented in the data.
- Insights gained can help optimize marketing strategies, improve customer engagement, and increase sales for Amazon sellers and marketers: The insights gained from the analysis can be used by Amazon and its sellers to optimize marketing strategies, improve customer engagement, and increase sales. For example, sellers can use the insights to optimize their product listings and pricing strategies, while Amazon can use the insights to improve its overall sales performance.
- Future insights could include identifying seasonal trends, improving supply chain management, and enhancing product listings to increase sales performance: While the analysis provided valuable insights, there is still room for further analysis. Future insights could include identifying seasonal trends in sales performance, improving supply chain management to optimize inventory levels, and enhancing product listings to increase sales performance.

The analysis yielded specific and actionable insights that can help businesses optimize their marketing and sales strategies, improve customer engagement, and ultimately increase sales performance. Some of the key insights include identifying popular categories and subcategories, analyzing sales performance, understanding pricing strategies, evaluating customer ratings, determining discount effectiveness, assessing the impact of product images, and analyzing the performance of product links. By leveraging these insights, businesses can refine their marketing and sales tactics, track sales performance, optimize pricing strategies and discounts, and enhance product design and marketing efforts to better cater to their target audience.

PART 3: DATA SOURCES

The dataset used consists of amazon sales data.This dataset was sourced from Kaggle.
It contains 9 columns, namely: Name, Product_category, Image, Link, Ratings, No_of_ratings, Discount_price, Actual_price.

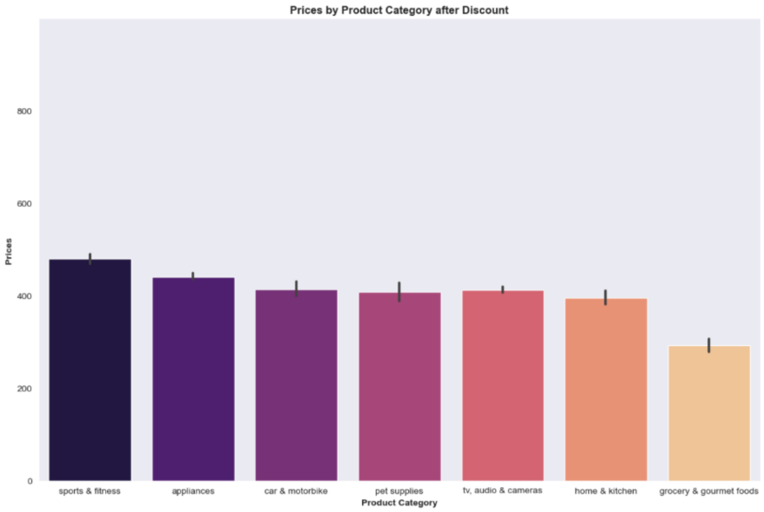
	Column Name	Description	Data Format	Validation
1.	name	The name of the product	String	Required
2.	product_category	The category that the product belongs to	String	Optional
3.	image	The URL of the product image	String	Required
4.	link	The URL of the product page	String	Required
5.	ratings	The average rating of the product (out of 5)	Float	Optional
6.	no_of_ratings	The number of ratings the product has received	Integer	Optional
7.	discount_price	The discounted price of the product	Float	Optional
8.	actual_price	The actual (original) price of the product	Float	Required
9.	discount_percentage	The percentage of discount on the product (if applicable)	Float	Optional

The dataset contained various worksheets based on product_category and was merged into a single combined dataset to help gain insights on the product_category and analysis on the trends related to marketing strategies.
The final dataset was obtained by
1) Checking and removing null values,
2) Checking and correcting incorrect values,
3) Dropping records which consisted of null values in columns which could not be cleaned further.

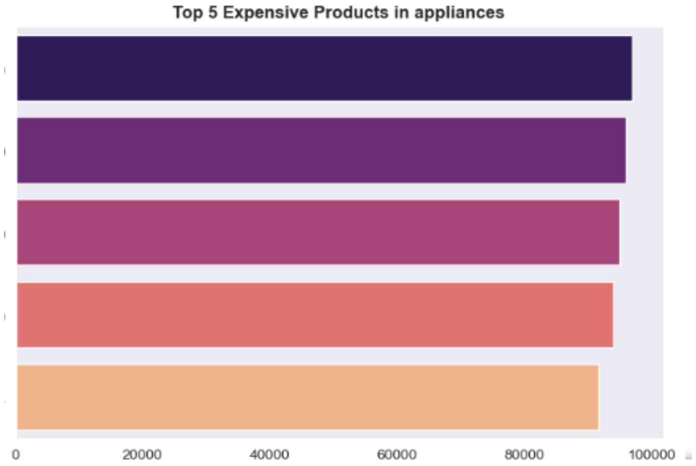
The dataset analyzed contained information on all product sales on Amazon. To investigate the relationship between the discount_price and actual_price, a new column called discount_percentage was added. This was done to better understand customer behavior in relation to this factor and how it correlates with the price difference before and after the discount.

PART 4: RESULTS AND METHODS

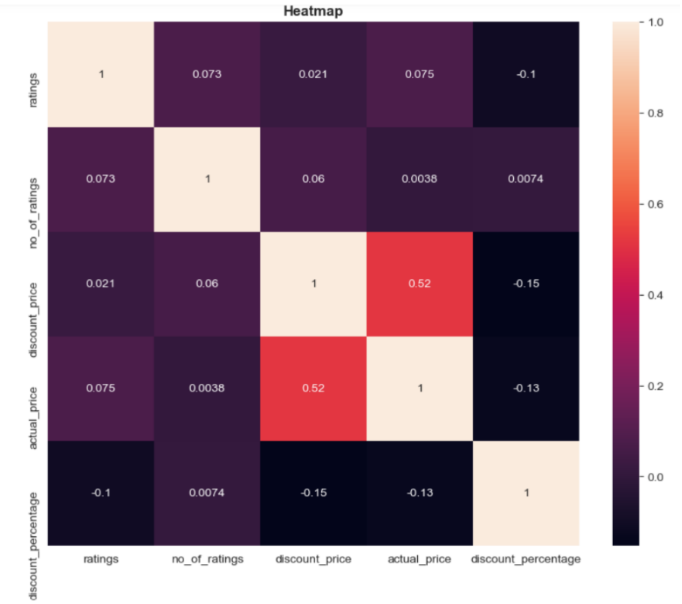
The dataset obtained from Kaggle had a significant number of null values in four columns, which required removal to obtain usable data for visualizations and dashboards. During the initial exploratory data analysis (EDA), we identified inconsistencies in the "ratings" column, with incorrect values in 105 records and 856 null values. We addressed this by using Python to remove the incorrect and null values and replacing them with the mean of the entire column. However, for the remaining null values in the "no_of_ratings," "actual_price," "discount_price," and "discount_percentage" columns, we decided that dropping the records was a better option to avoid manipulating data and producing inaccurate results. This left us with 23,382 records for further analysis.



Using the cleaned dataset and the seaborn library in Python, we performed various visualizations to gain insights into the data. We plotted a bar graph to compare product categories and their discounted prices and found that even after the discount, the "Sports & Fitness" category had a higher price range.

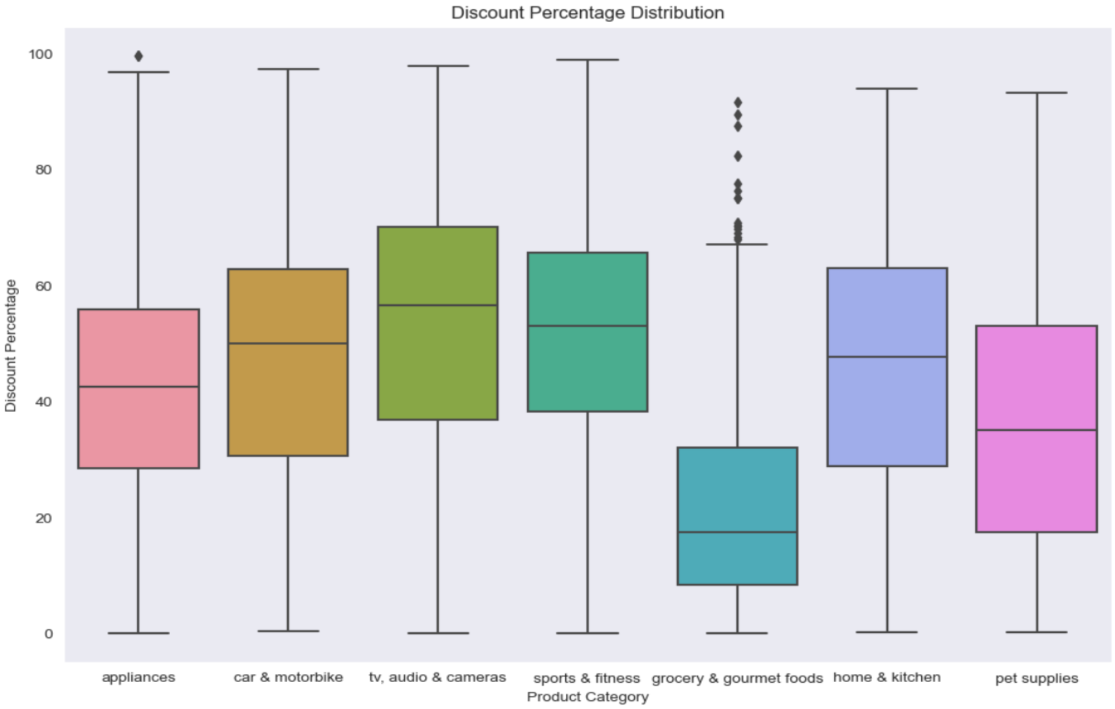


We also plotted graphs to identify the top five most expensive and cheapest products in each category, with "Appliances" and "TV, Audios & Cameras" having the most expensive products, while all the cheapest products belonged to the "Grocery & Gourmet Foods" category.



We then created a heatmap to understand the correlation between columns and found that "actual_price" and "discount_price" had a positive correlation of 0.52.

Using a scatterplot, we observed that the higher the actual price, the lower the discount, and the lower the actual price, the higher the discount. We gained insights that many products in our dataset had prices ranging from 0-5000. Additionally, we plotted a line graph to compare ratings and the number of ratings for each category and found that the "TV, Audio & Cameras" category had the highest number of products and an average rating of 4-4.5.

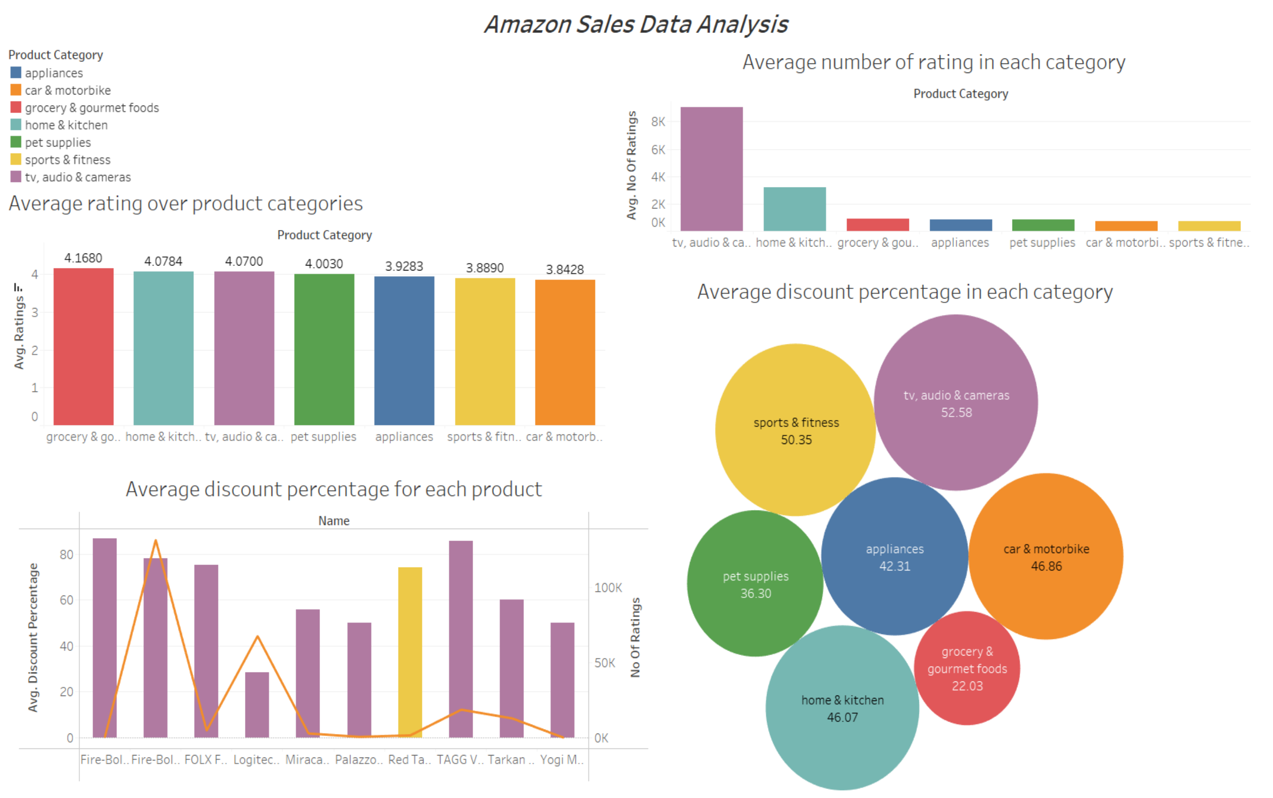


To visualize the distribution of discount percentages, we created a whisker plot/box plot, and observed that the median for the "Grocery & Gourmet Foods" category was closer to the lower quartile, indicating a lower distribution in this range. However, there were many outliers, indicating a larger distribution range for this category. We also noted that the discount percentage for many products was in the range of 50-65%.

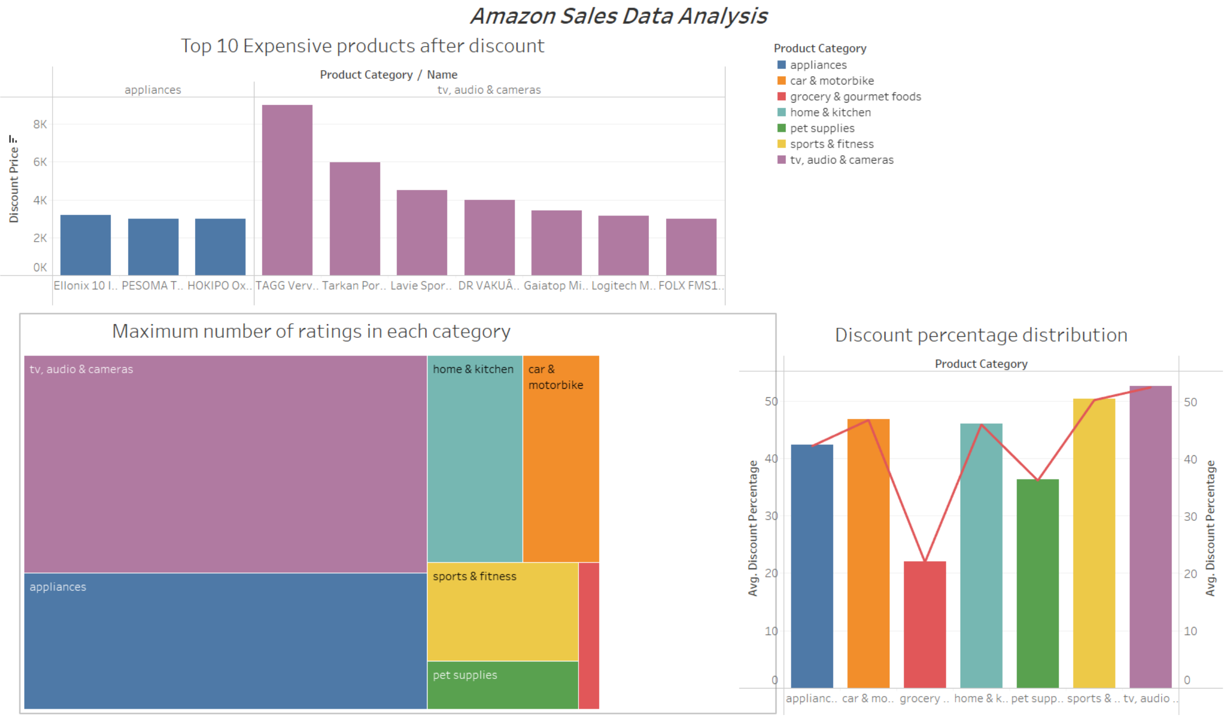
After exporting the cleaned dataset to a CSV file and importing it into Tableau, we created two dashboards to serve as a tool for apartment hunting. We developed two dashboards to help users gain valuable insights for making informed purchasing decisions. The first dashboard provided a high-level overview of the relation between rating and discount percentage for each product category, while the second dashboard allowed for a direct comparison between them. Our aim was to ensure that both dashboards were easily comprehensible and accessible to the end user.

Dashboard 1

Our first dashboard was designed to provide Amazon with a comprehensive view of the sales performance of each product category and to facilitate a comparative analysis between categories. Viewers can easily filter and select a specific category to view the average ratings, discount prices, and discount percentages for that category. The dashboard also provides information on the average ratings and discount percentages across all categories to help gauge overall customer satisfaction and the effectiveness of discounts. This allows Amazon to make informed decisions about marketing strategies, inventory management, and product development to optimize sales performance and customer satisfaction.



Dashboard 2



The second dashboard in our analysis provided valuable insights into the top 10 most expensive products and their respective categories. Additionally, we were able to visualize the distribution of maximum ratings in each category and the distribution of discount percentages across categories, providing a comprehensive understanding of how pricing and discounts affect product performance in each category. These insights can be leveraged by businesses to optimize their pricing strategies and improve their product offerings.

Overall, our project yielded significant insights from the exploratory data analysis and visualization. Our findings provided actionable insights that can help our target audience make informed decisions about their purchases on Amazon.

PART 5: LIMITATIONS AND FUTURE WORK

We studied the trends and patterns of the Amazon sales to draw data-informed business insights from our wrangled dataset. According to our analysis we made some recommendations to Amazon to further improve and enhance their sales and revenue. We inferred that Tv, Audio, Electronics and other appliances were the strong suites of Amazon generating the most revenue and having the highest user engagement. We were able to derive this from the number of ratings (400k+) and the highest average rating of 4.1/5. This stood as a clear indication that this category was the fan favourite. Our analysis further dove into studying the interrelationship between user engagement and pricing strategy. Hence, to further bolster our derivation we found out that this category was also given the maximum average discount percentage amongst all others which in turn resulted in higher CSAT (customer satisfaction) ratings. Based on these trends, we drew more insights from the data and recommended Amazon to invest in marketing of the groceries segment as the CSAT ratings were high, but the number of customers were low. As against that, for automobile parts, the demand was high, but CSAT was low. This indicated the need for improvement in those products.

The risks and caveats for the recommendations arise in the estimation of buyers. Our dataset comprises of the number of ratings given by the customers and not sales completed. We assume that a user that submits a complete review is termed as a 'Power user' and he is a customer of the product. To mitigate this risk, we have cleaned all the incomplete reviews and null values from our dataset to bring our assumptions closer to actuality.

A future scope for this study can comprise of pulling actual sales data along with their reviews to strengthen the claims.

- We further plan to retrieve the sales made along with their dates. The distribution of this data will allow us to study the performance trends of the categories according to seasons and time of the year.
- Our current database is pulled from Amazon sales in India hence the prices are in INR. We further plan to analyze the data over larger demographics consisting of multiple countries to study geographical variations in product sales.
- A wide aspect data analysis can be conducted to study the Covid-19 effect of each product category. This can be done by including the sales datasets from 2018 to 2023.