

Nikhil Sharma

Customer Shopping Data

17 July 2023

# Data Analysis Report: Amazon Sales Data

## 1. Introduction

This report presents findings from the analysis of the "Amazon Sales Data" dataset. The dataset, sourced from Kaggle.com (uploaded by Lokesh Parab), provides information about online sales in the USA. It includes data on various products, merchandise, and electronics sold in different states. The goal of this analysis is to gain insights into sales trends, customer behavior, and pricing patterns to assist large retailers in optimizing their strategies and increasing sales.

## 2. Data Preparation

The necessary libraries, including numpy, pandas, matplotlib.pyplot, seaborn, os, and glob, were imported to facilitate data analysis. The dataset was loaded into a Pandas DataFrame by merging multiple CSV files from different categories. Irrelevant columns, such as "Unnamed: 0", "image", and "link", were dropped to focus on the relevant variables.

## 3. Data Exploration and Cleaning

The dataset was examined to understand its structure and content. It consists of 1,103,170 rows and 10 columns. The column names were inspected to identify the available variables for analysis. Missing values were identified in several columns, which were addressed in subsequent steps.

## 4. Data Type Conversion

To facilitate analysis, appropriate data types were assigned to relevant columns. The columns "ratings", "no\_of\_ratings", "discount\_price", and "actual\_price" were converted to numeric data types using the appropriate functions. String characters and formatting symbols were removed before conversion.

## 5. Handling Missing Values

Missing values were observed in columns such as "ratings", "no\_of\_ratings", "discount\_price", and "actual\_price". The missing values in the "ratings" and "no\_of\_ratings" columns were imputed using the median values within each sub-category and main category. Rows with

missing values in both the "discount\_price" and "actual\_price" columns were dropped from the dataset.

## 6. Exploratory Data Analysis

The analysis focused on understanding the distribution and relationships within the dataset. The number of main categories and sub-categories was determined, providing insights into the product variety. The minimum and maximum prices for both actual and discounted prices were calculated, revealing price ranges across different categories.

### 6.1 Sales Trends:

- **Main Categories:** Accessories is the highest selling at \$222,342, ahead of Women's Clothing and Men's Clothing.
- **Sub-Categories:** Men's Fashion topped the list at \$38,088, with Clothing and Western Wear next.

## 7. Ratings and Reviews Analysis

The dataset was analyzed to determine the average ratings and number of ratings received for each main category and sub-category. This analysis highlighted the categories with the highest and lowest ratings, as well as those with the most and least number of ratings. The results were visualized using bar plots to facilitate interpretation.

- **Main Categories:** Grocery & Gourmet Foods had the highest mean rating of 4.06, ahead of Toys & Baby Products and Pet Supplies.
- **Sub Categories:** International Toy Store topped the list at 4.57, with Value Bazaar and All Grocery & Gourmet Foods next.
- Accessories had the highest combined total rating of 866,624 and Bags & Luggage in the sub-category at 152,079.
- On average, people were more likely to review grocery & gourmet foods, home & kitchen, and stores and Value Bazaar for sub-category.
- TV, audio & cameras had the most number of reviews and All Electronics in sub-category.
- Most ratings were between 3.5 and 4.2.

## 8. Price Analysis

The average discounted price and actual price were analyzed by main category and sub-category. This analysis provided insights into pricing strategies and variations across different categories. The findings were visualized using bar plots and logarithmic scale plots to emphasize differences.

- **Main Categories:** Accessories and Appliances offered the highest average discounted price.
- **Sub Categories:** Air Conditioners and Televisions offered the highest average discounted price.

## 9. Conclusion

Based on these findings, retailers can make informed decisions to optimize their strategies and increase sales:

- Focus on promoting and expanding the Accessories category, as it shows strong sales performance.
- Pay attention to customer reviews and ratings, especially in high-rated categories like Grocery & Gourmet Foods, to maintain positive customer sentiment and enhance reputation.
- Consider pricing strategies in line with market trends, offering competitive prices and discounts, particularly in categories such as Accessories and Appliances.
- Explore opportunities for expansion and improvement in sub-categories that show high potential, such as Men's Fashion and International Toy Store.

## 10. Recommendations

Based on the analysis, the following is recommended:

- **Investigate the relationship between product categories and customer preferences:** Analyze customer behavior within different product categories to identify patterns and preferences. This can help retailers tailor their product offerings and marketing strategies to specific customer segments.
- **Conduct market research to understand customer preferences and trends in different shopping malls:** Gather data on customer preferences, shopping habits, and trends in various shopping malls or locations. This information can provide insights into regional variations and help retailers customize their offerings accordingly.
- **Implement targeted marketing campaigns for specific customer segments:** Utilize the findings from customer demographics and shopping habits analysis to develop targeted marketing campaigns. By understanding the preferences and behaviors of different customer segments, retailers can create personalized and effective marketing strategies.
- **Explore the impact of pricing strategies on customer purchasing behavior within different product categories:** Analyze the relationship between pricing strategies, customer purchasing behavior, and revenue generation. Experiment with different pricing models and discounts to identify the most effective approach for each product category.

## 11. Acknowledgments

I would like to express our gratitude to Lokesh Parab, who uploaded the "Amazon Sales Data" dataset on Kaggle, which served as the foundation for this analysis. We acknowledge the efforts and contribution of Lokesh Parab in making this dataset publicly available for research purposes.

## 12. References

The analysis used the following libraries and tools:

- NumPy: A library for numerical computing in Python.
- Pandas: A data manipulation and analysis library.
- Matplotlib: A plotting library.
- Seaborn: A data visualization library.
- Jupyter Notebook: An interactive coding environment.

For a detailed understanding of the analysis process, including the Jupyter Notebook code used, please refer to the original analysis documentation.

Best regards,

Nikhil Sharma