PRODUCT RECOMMENDANDATION AND SENTIMENT ANALYSIS

CONTENT

01 INTRODUCTION

02 PROBLEM STATEMENT

03 BUSINESS UNDERSTANDING

04 DATA UNDERSTANDING & PREPARATION

05 DATA ANALYSIS

06 MODELLING

07 CONCLUSION

INTRODUCTION

E-commerce in Kenya has experienced remarkable growth, with platforms like Jumia drawing a significant user base. However, navigating vast and often poorly organized product catalogs poses challenges for users. Additionally, the lack of robust recommendation systems and unreliable reviews diminishes the shopping experience, leading to decreased user satisfaction.

To address these gaps, we propose a recommendation and sentiment analysis system that harnesses user interaction data and feedback. This solution aims to deliver tailored product suggestions and actionable customer insights, enhancing user satisfaction.

PROBLEM STATEMENT

Inadequate product recommendations based on user behavior

Unreliable rating and review system

Limited retailer insights into customer sentiment

BUSINESS UNDERSTANDING

Our target audience are:

- Primary Users (Consumers): Customers who frequently shop on Jumia and need relevant recommendations to streamline their product search and improve purchase decisions.
- Retailers/Sellers: Businesses and individual sellers on Jumia who seek insights into customer preferences and sentiment to better tailor their product offerings and marketing strategies.

DATA UNDERSTANDING

For this project, we used a dataset we scrapped from Jumia. The dataset includes product reviews and ratings, which are essential for building both the sentiment analysis and recommendation system models.

The relevant features, such as ‘user name’, ‘product name’, ‘category’, ‘review’ and ‘ratings’, are well-defined to facilitate both sentiment classification and recommendation tasks.

DATA PREPARATION

Handling missing values

Identify and address gaps in the data by imputing missing values and removing incomplete records to ensure consistency

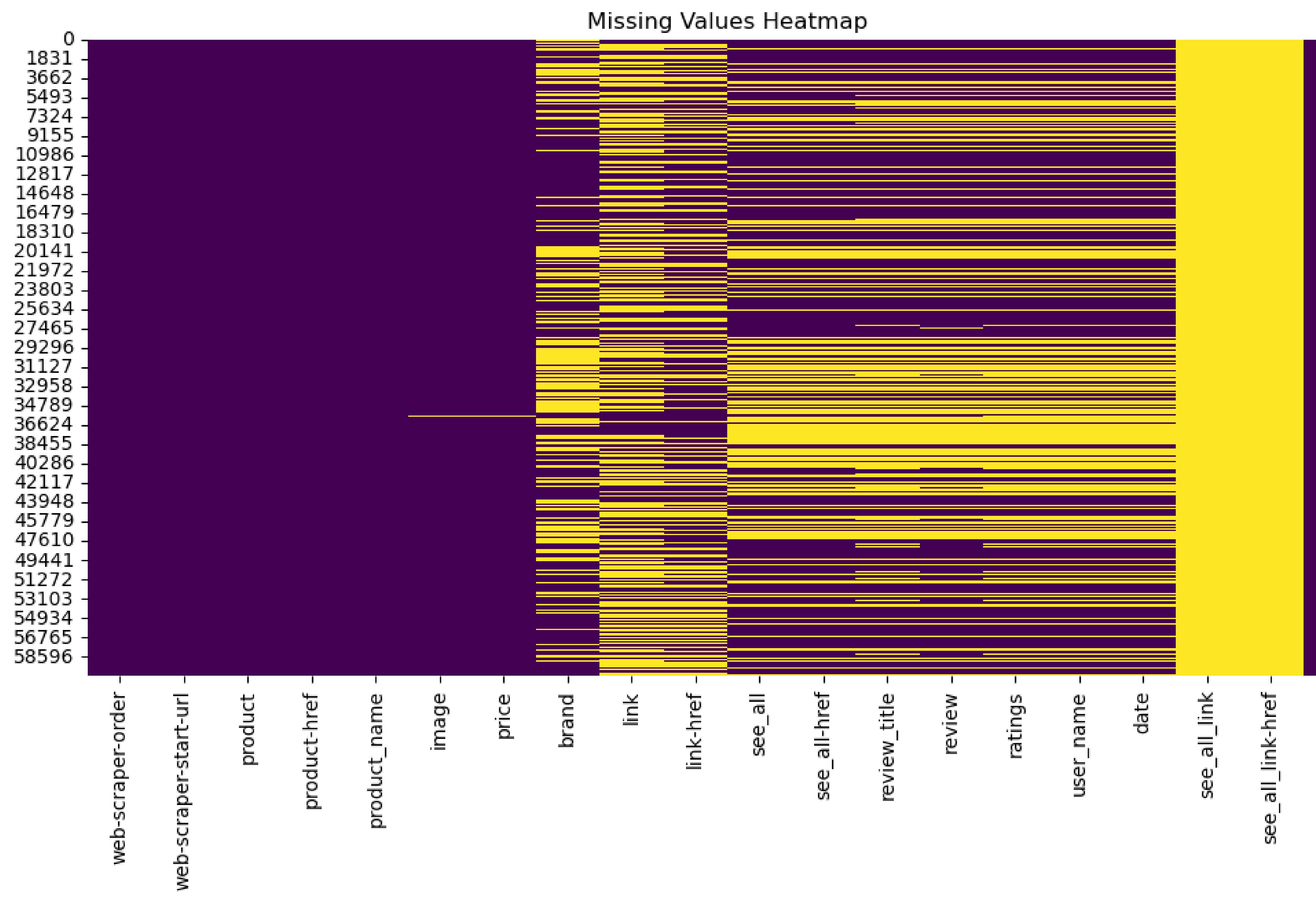
Feature selction & engineering

Select the most relevant variables and created new ones to improve model performance and highlight underlying patterns

Standardization

Convert categorical variables into binary columns and scale numerical features to a uniform range for better compatibility with our machine learning algorithms.

DATA CLEANING


Missing Values Heatmap

This heatmap provides insight to the following:

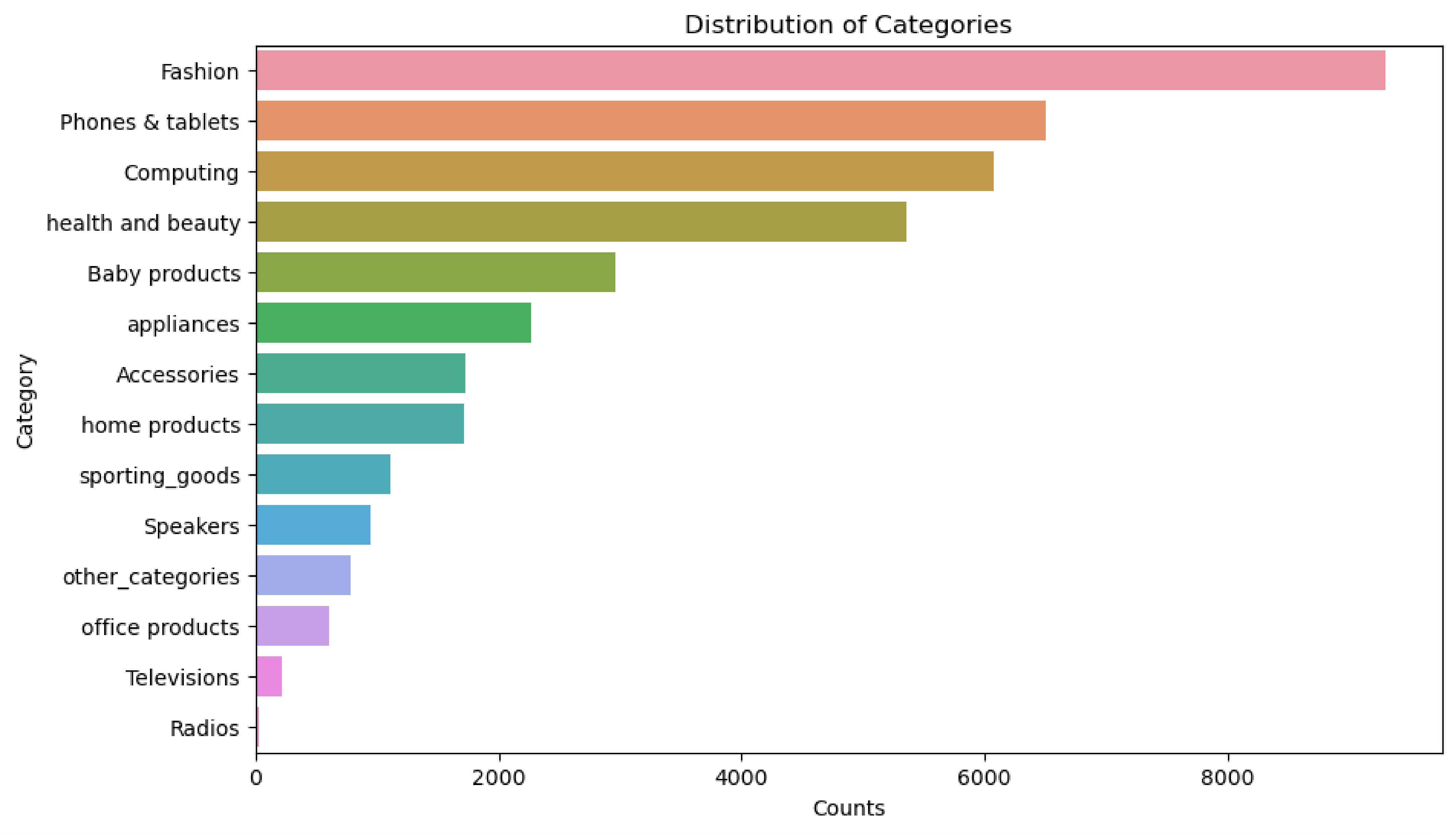
- Which columns need data cleaning
- Columns that might need to be dropped or imputed
- Potential systematic issues in data collection

DATA ANALYSIS

CATEGORIES

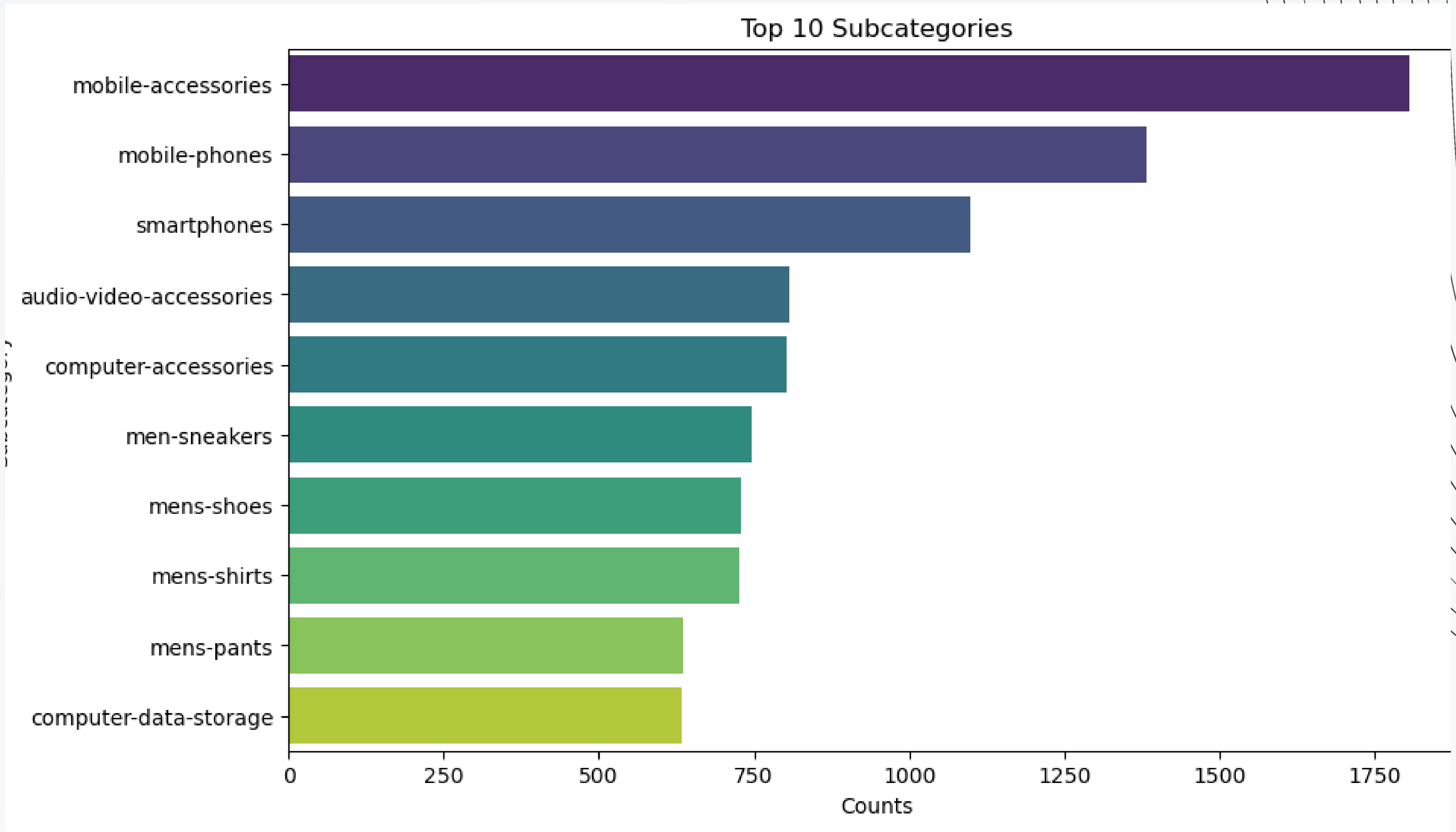
The distribution of categories provide the following insights:

- *Customer Interest*: The top categories reflect higher customer interest or demand.
- *Inventory Planning*: Retailers could focus more on stocking and promoting items in the most frequent categories.
- *Recommendation System Focus*: A recommendation system might prioritize building recommendations for these popular categories



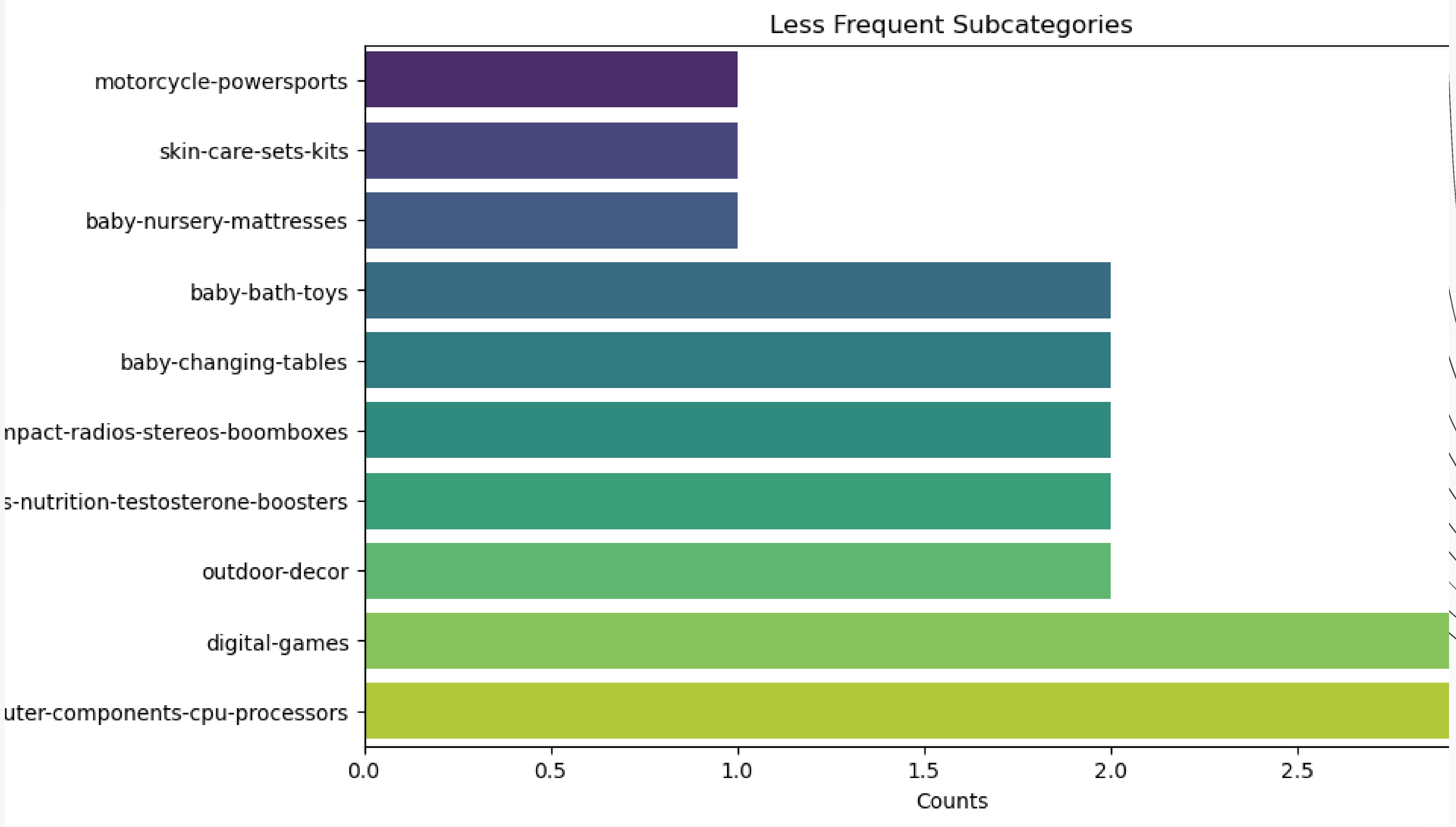
SUBCATEGORIES

The top subcategories represent popular product types, high customer demand, or a focus on certain items.



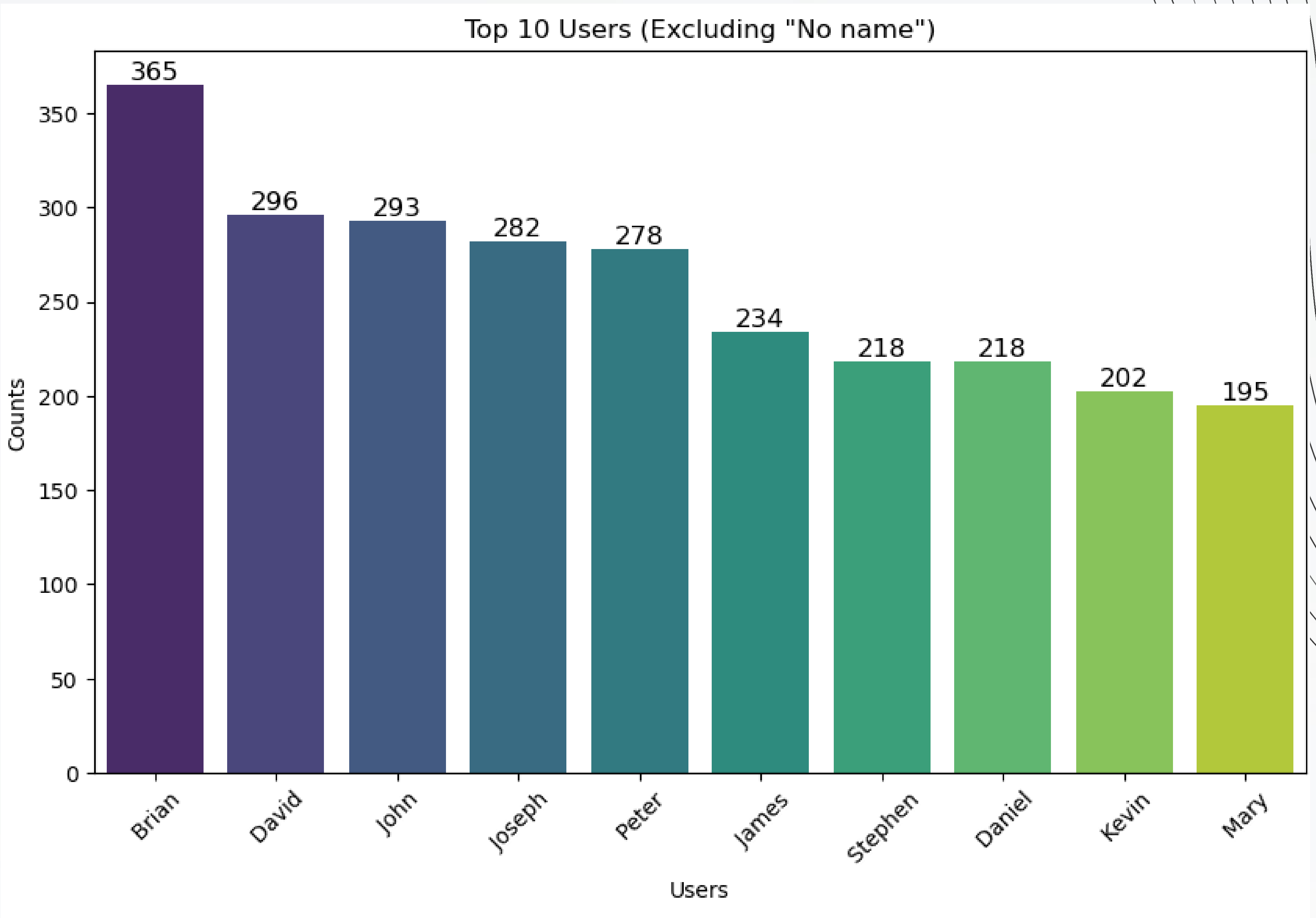
SUBCATEGORIES

The plot shows less frequent sub-categories potentially indicating lower popularity or inventory for these items.



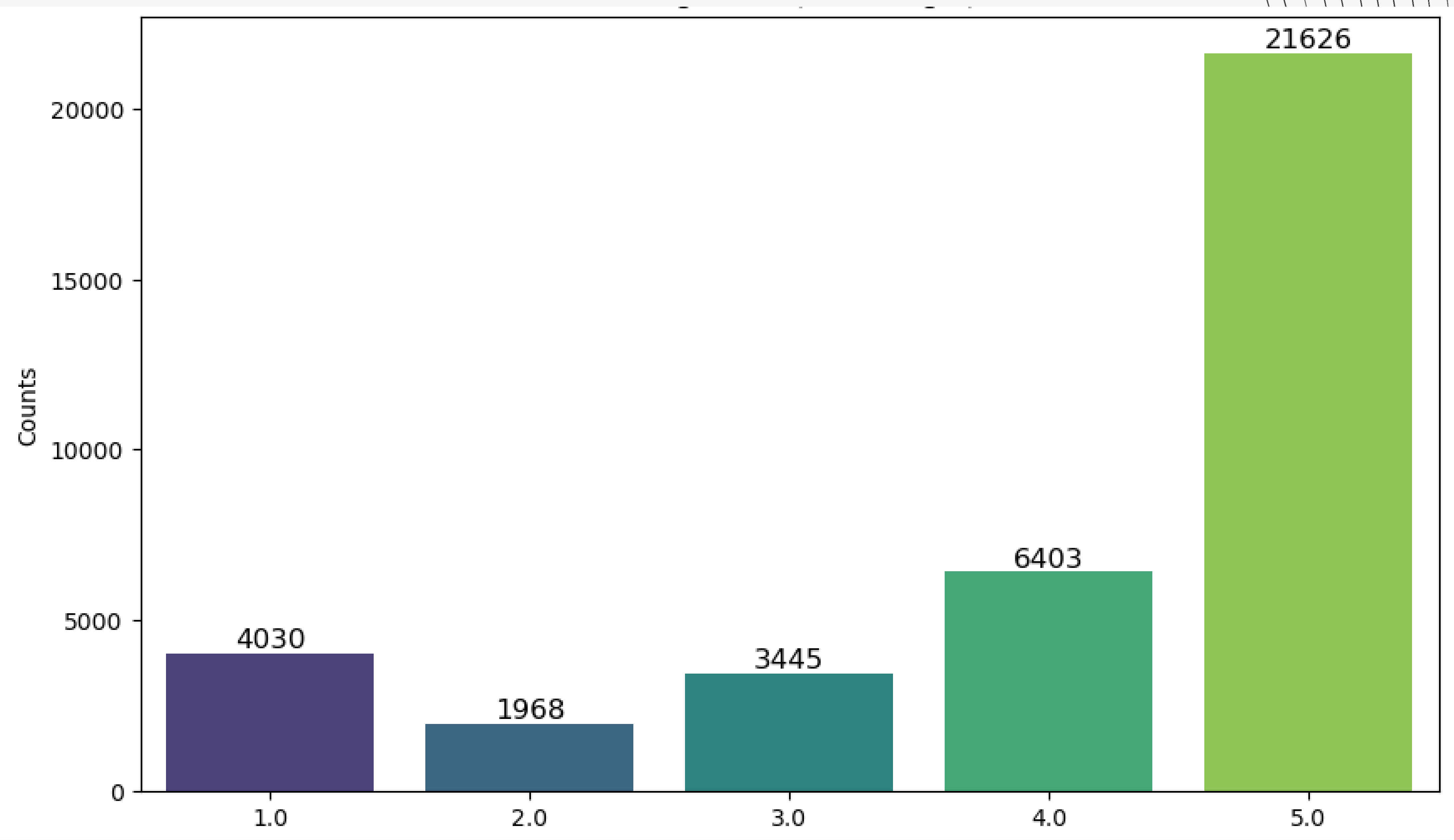
USERS

This distribution suggests that a few users are significantly more engaged than others in providing feedback. These top users are repeat customers or highly active reviewers who frequently purchase and leave feedback on products.



RATING COUNT

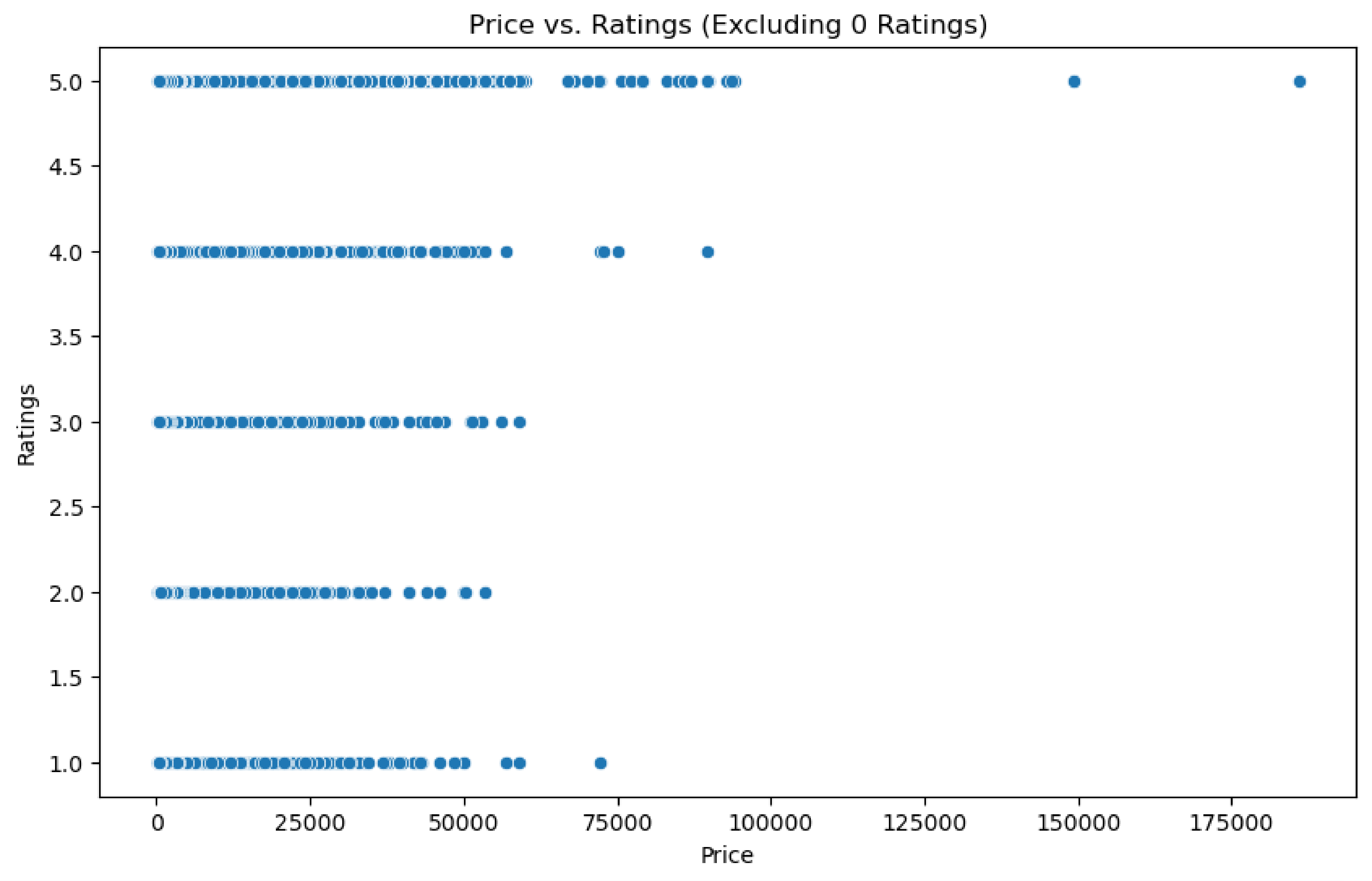
From the rating count plot, we can see that 5 is the most rating used by the website users followed by 4 which is a clear indication majority of the products are satisfying and pleasing to the buyers.



PRICE VS RATING

The scatter plot shows that:

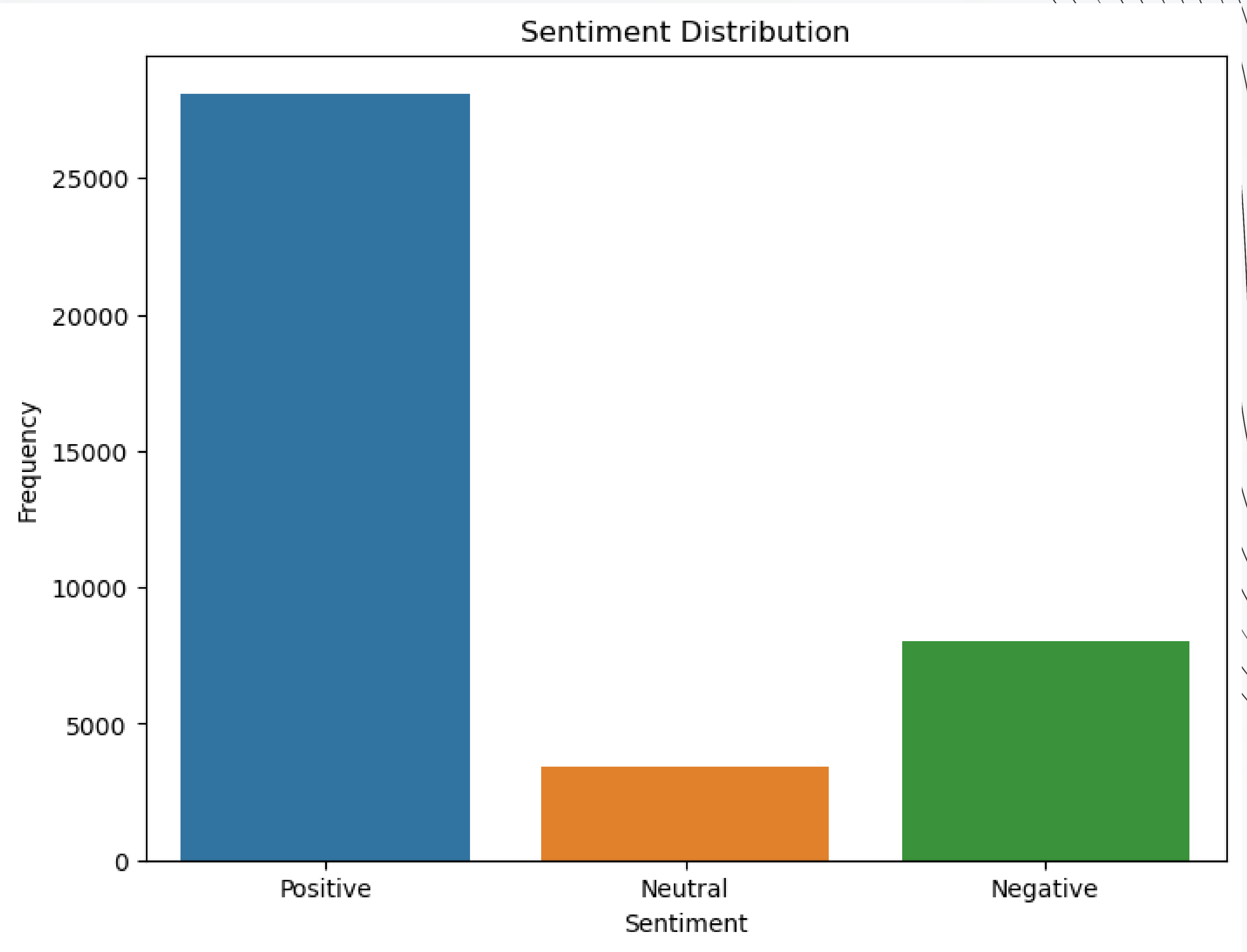
- Products have a broad range of prices for each rating level.
- Higher ratings (3 to 5) are more frequent than lower ratings (1 or 2).
- There’s no direct correlation between price and rating based on this visual.



SENTIMENT ANALYSIS

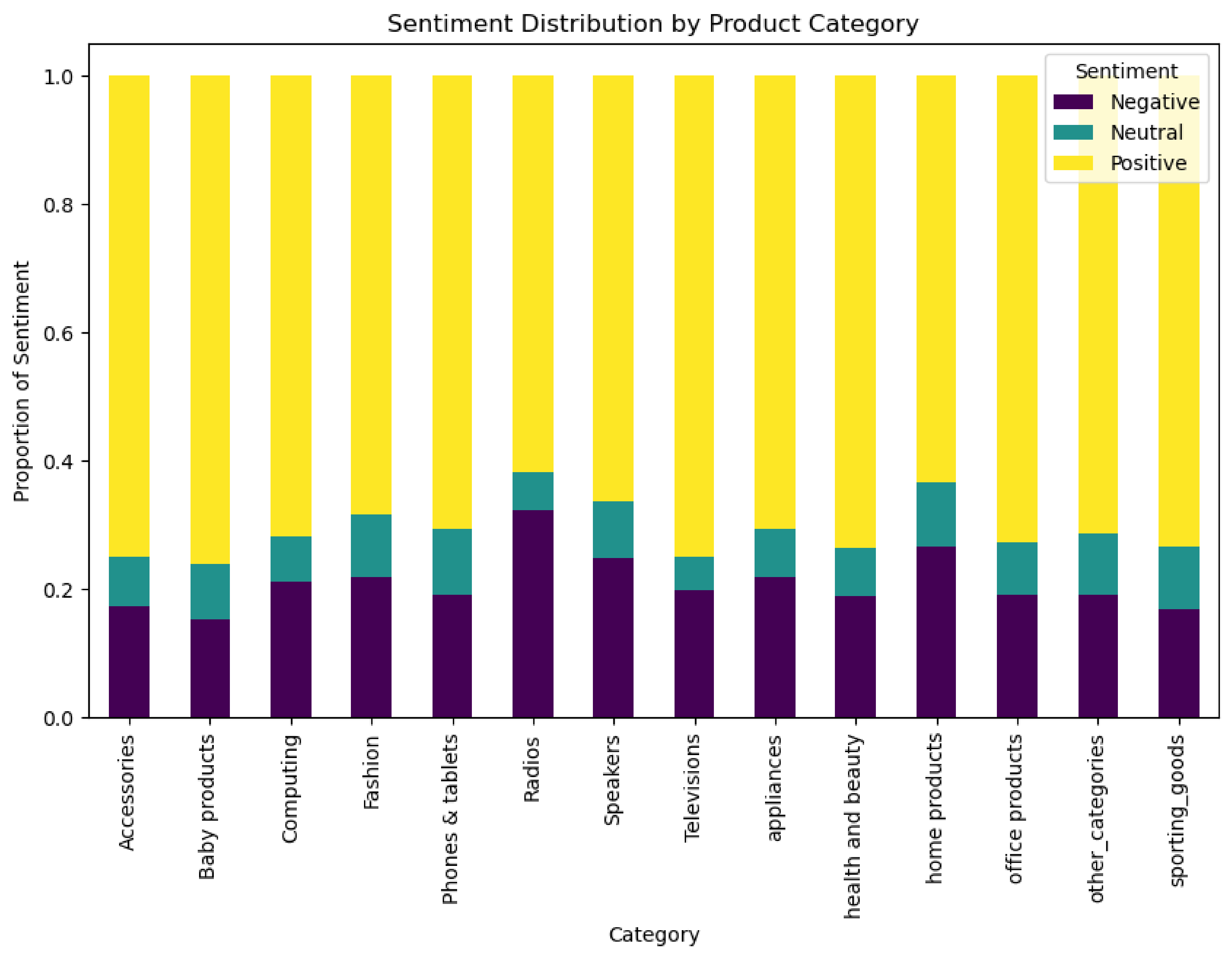
SENTIMENT DISTRIBUTION

- The distribution reveals that positive sentiments are dominant, with a significantly higher frequency compared to neutral and negative sentiments.



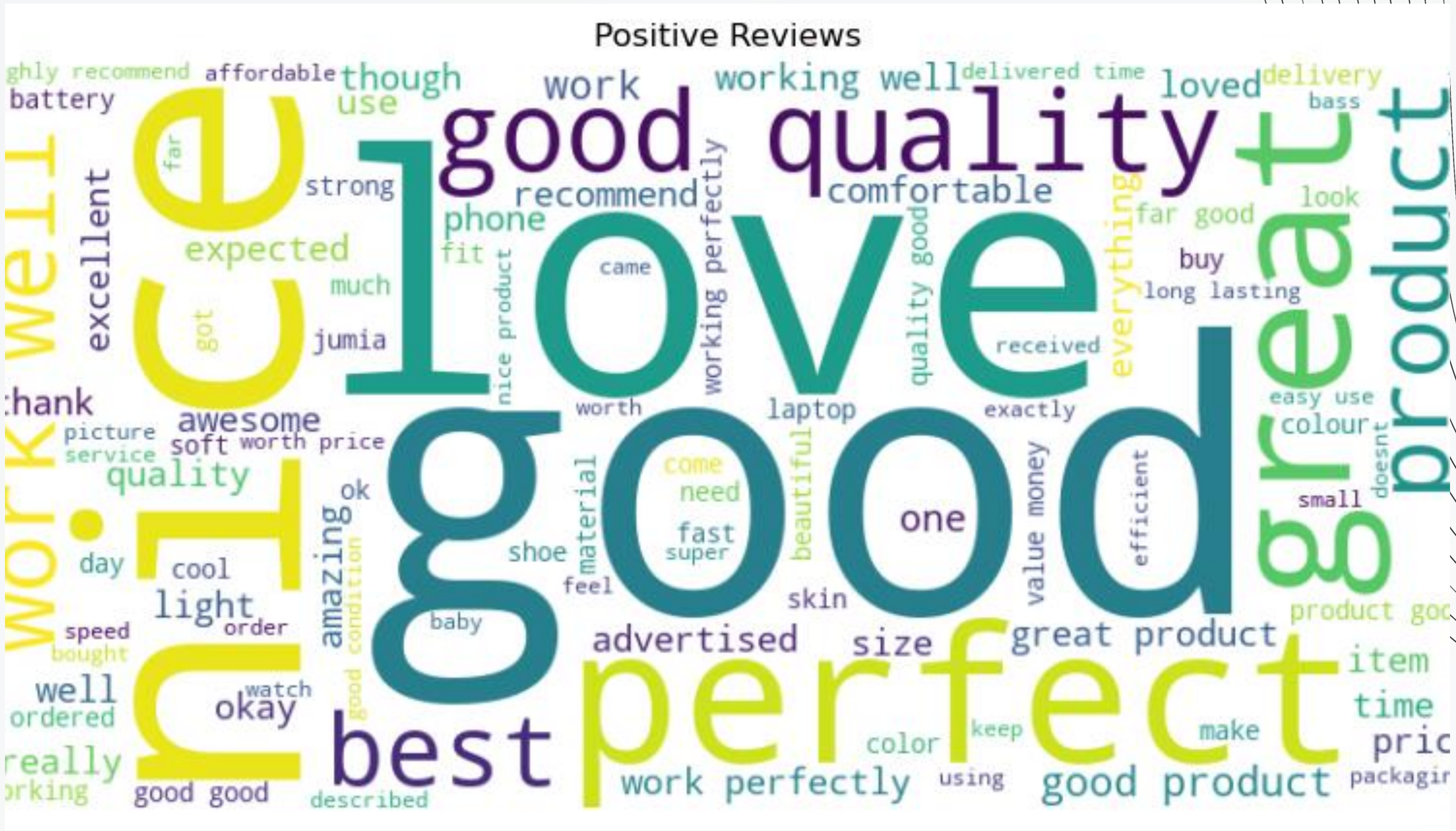
SENTIMENT BY CATEGORY

- Positive sentiment dominates across all product categories.
- Neutral and negative sentiments have smaller shares, with slight variations between categories.
- Some categories (e.g., Phones & Tablets and Televisions) seem to have a slightly higher proportion of neutral or negative sentiments compared to others.



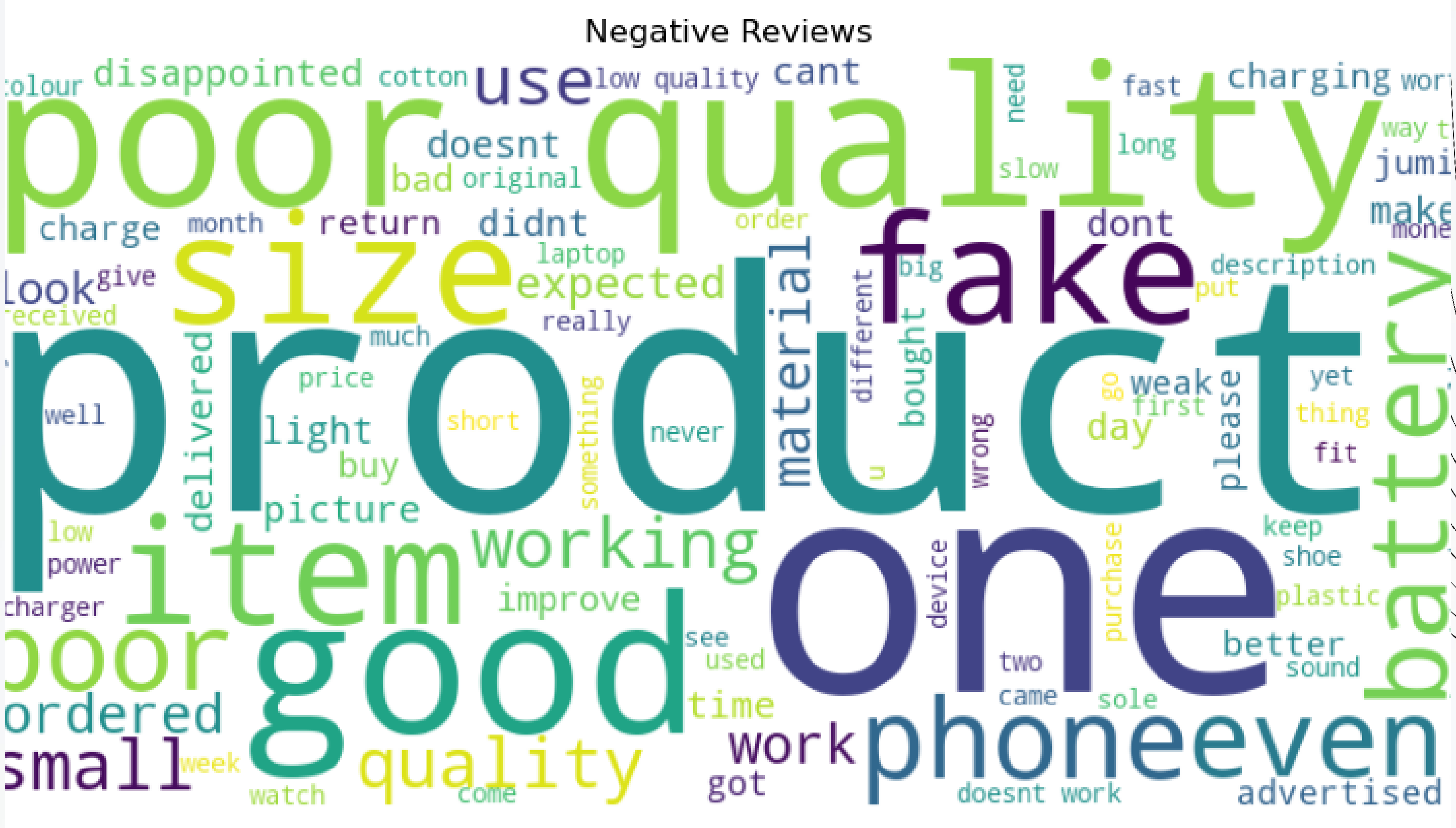
POSITIVE REVIEWS

The following words represented satisfaction and high ratings from users



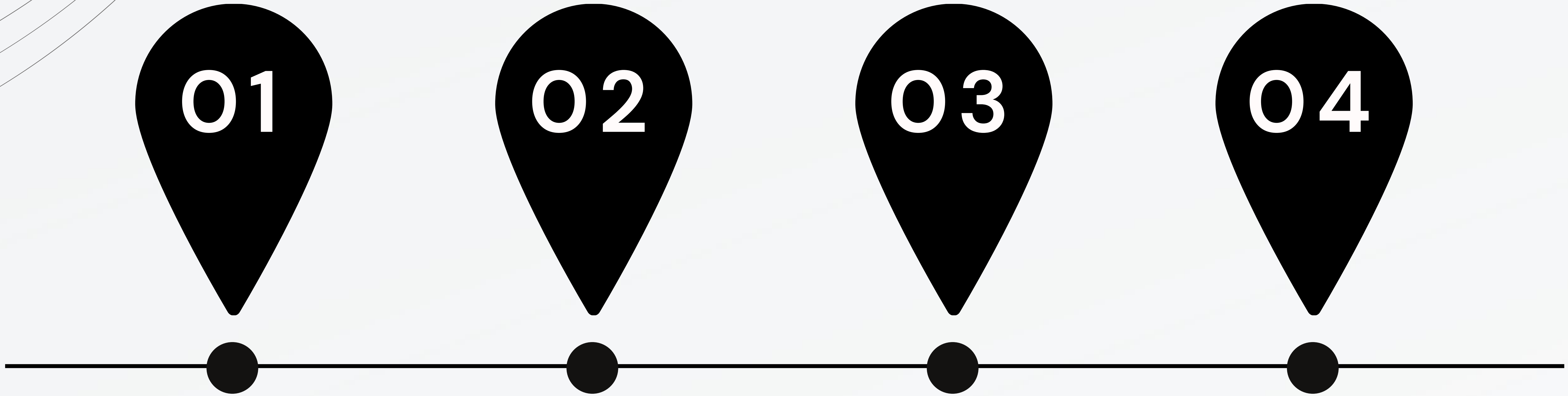
NEGATIVE REVIEWS

The following words represented poor ratings and dissatisfaction from users.



Negative Reviews

MODELLING



01

BOW

The model performs well on Positive sentiment but struggles with Neutral sentiment, evident from the low recall (0.30) and F1-score (0.37).

02

TF-IDF

Both ‘Positive’ and ‘Negative’ sentiment scores remain strong, with relatively high precision and recall.

03

NAIVE BAYES

Naive Bayes achieves a higher accuracy (0.82) compared to earlier models (0.75–0.77). This model assumes feature independence, which works well for distinct patterns (e.g., clear Positive and Negative keywords).

04

RANDOM FOREST

The random forest classifier performs well for "Positive" and "Negative" sentiments.

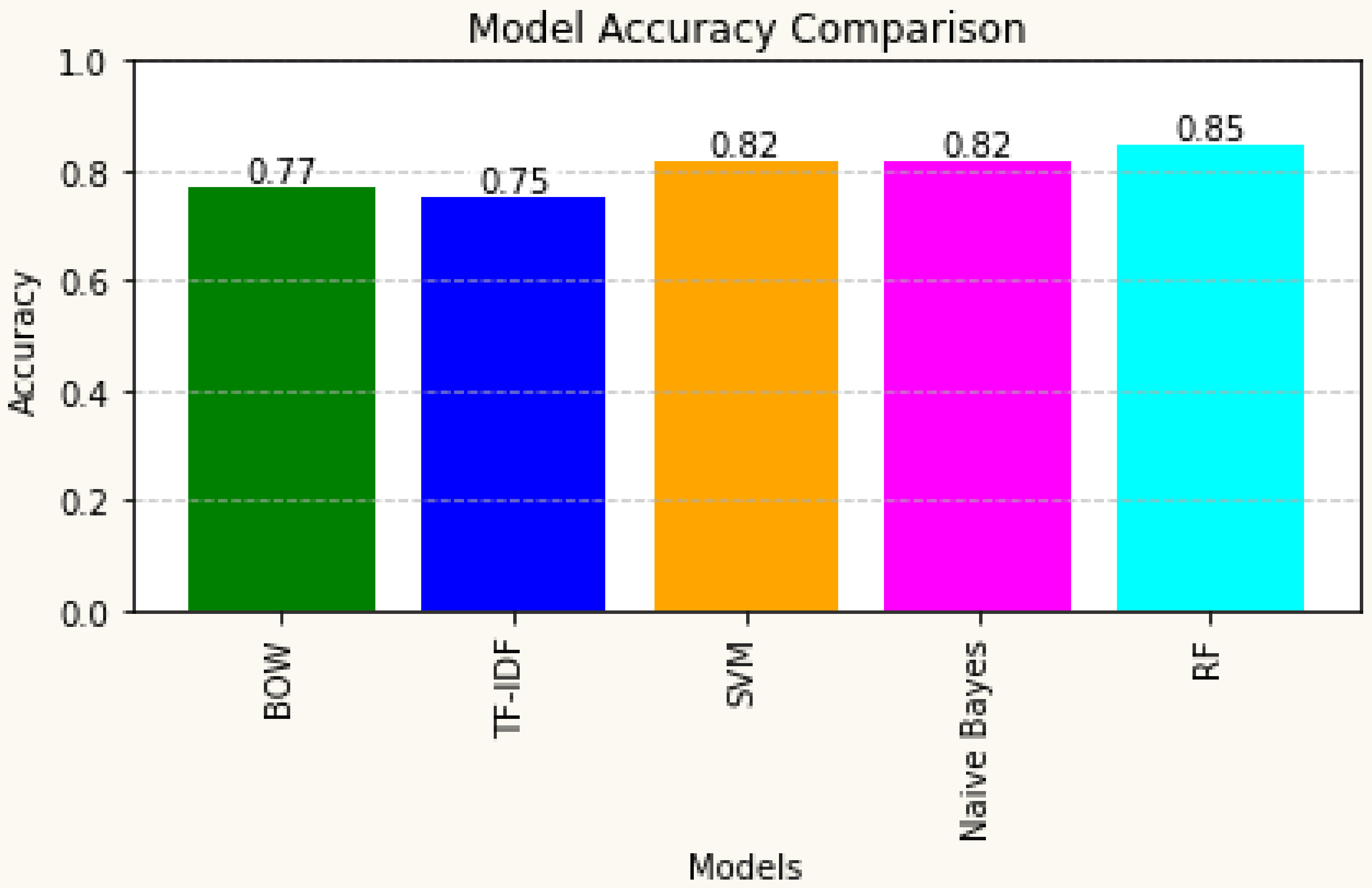
MODEL COMPARISON

* More sophisticated models (RF, SVM, Naive Bayes) outperform simpler approaches (BOW, TF-IDF)

* Random Forest leads with 85% accuracy, suggesting it's best at capturing complex sentiment patterns

* The performance difference between SVM and Naive Bayes is negligible (both 82%)

* All models achieve above 75% accuracy, indicating good baseline performance for sentiment analysis



RECOMMENDATION SYSTEM

Unpersonalized Recommendation

For this, we recommended items and categories with the highest weighted ratings.

Item-Based Recommendation

Recommends items similar to those the user has liked in the past.

Content-Based Filtering

Uses item metadata (e.g., product descriptions, tags) to recommend similar items to those the user has previously liked.

RESULTS

1. **Sentiment Analysis:**
    - Best Model: Random Forest achieved the highest performance across evaluation metrics, showcasing its effectiveness in classifying customer sentiments.
2. **Recommendation Systems:**
    - *Unpersonalized System:* Delivered general recommendations based on popular and highly rated items, effective for broad user bases.
    - *Item-Based Collaborative Filtering:* Successfully suggested products by analyzing similarities between items that users interacted with.
    - *Content-Based Filtering*: Accurately provided personalized recommendations by analyzing product attributes and user preferences.

CHALLENGES

1. *Dataset Limitations:* Scraping the dataset revealed a restriction of 10 reviews per item, limiting the depth of data available for analysis.
2. *User Identification Issues:* The user_name field primarily contained first names, making it challenging to build a robust user-based recommendation system due to the lack of unique identifiers.

- *Impact:* These challenges required adapting our approach, such as focusing on item-based and content-based recommendation systems and leveraging available data effectively.

CONCLUSION

This project successfully developed and evaluated a dual-purpose system integrating product recommendation and sentiment analysis to address key challenges in Kenya's e-commerce landscape. By focusing on user interaction data and review analysis, the system provides actionable insights that enhance user satisfaction and retailer efficiency.

NEXT STEPS

1. *Improving Sentiment Analysis:*

- Use more advanced transformer models like BERT for higher accuracy.
- Add multilingual support for diverse user demographics.

2. *Expanding Data Scope:*

- Collect more diverse datasets, including competitor data or customer service logs.
- Analyze data trends over a longer time horizon for better forecasting.