DS5110 Section 4
Daniela Alejandra Gonzalez

Project: Customer Feedback Analysis and Summarization
Report on Sentiment Analysis and Percentage Computation for Amazon Reviews

Objective

This analysis processes Amazon reviews using PySpark to explore sentiment trends across different product categories. The primary goal is to understand the distribution of positive, negative, and neutral sentiments and compute their percentage contribution within each category.

Data Cleaning and Preprocessing

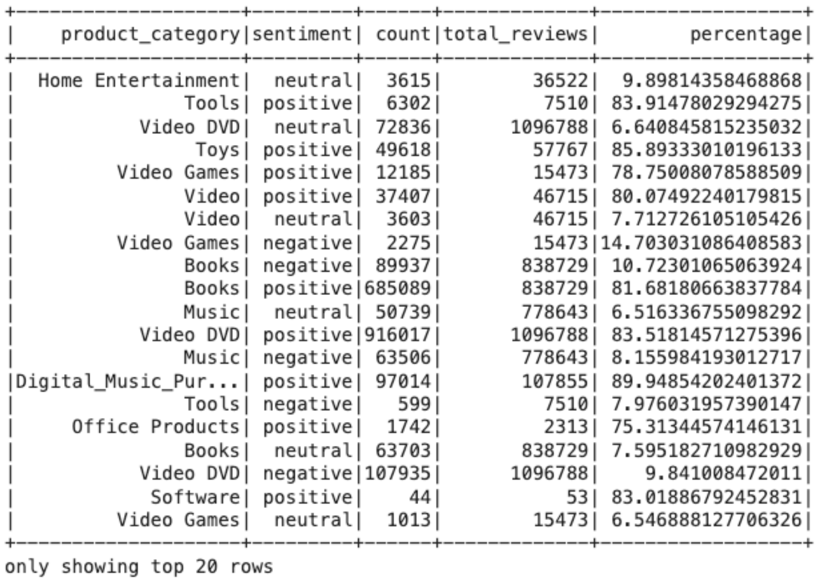
The dataset was cleaned by removing null or empty values in the `review_body` column to ensure data reliability. Text preprocessing steps included converting the text to lowercase, tokenizing the reviews into individual words, and removing stop words. These steps ensured a consistent format for sentiment analysis and helped focus on meaningful content.

Sentiment Analysis

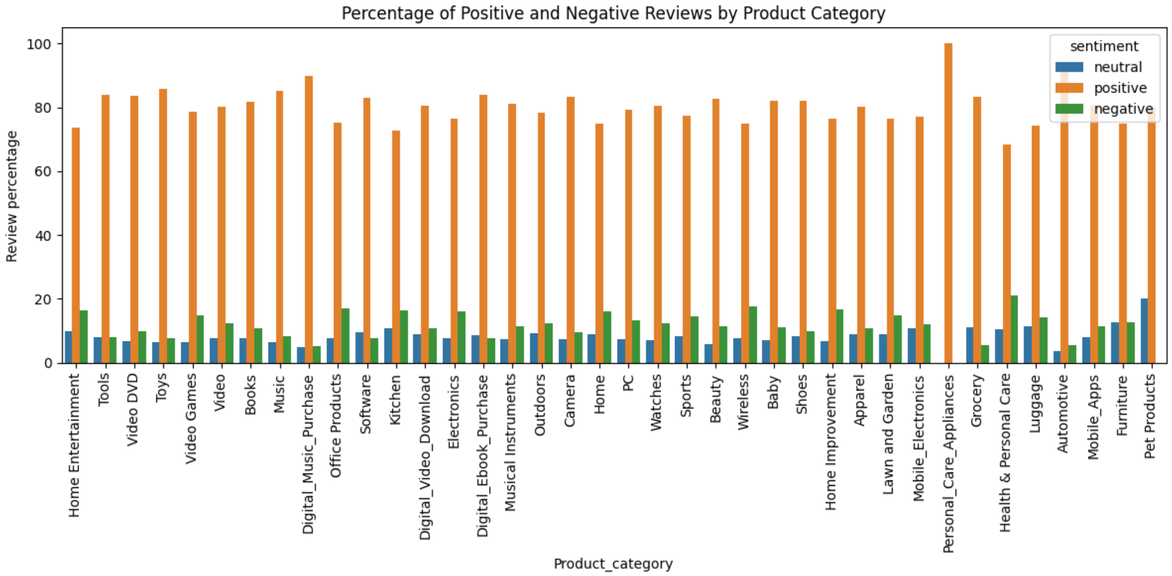
Reviews were categorized into sentiments based on their star ratings. Positive sentiment included reviews with 4 or 5 stars, negative sentiment comprised 1- or 2-star reviews, and neutral sentiment represented reviews with 3 stars. This labeling allowed a clear distinction between the sentiment types. Counts for each sentiment type were determined, providing a foundation for analyzing trends across product categories.

Product Category Analysis

The analysis grouped reviews by product category and sentiment, calculating the total number of reviews and the sentiment-specific counts for each category. To provide a deeper understanding, the percentage of reviews for each sentiment within a category was computed. For example, in the "Tools" category, 83.91% of reviews were positive, indicating high customer satisfaction. In contrast, categories such as "Books" and "Video Games" showed a higher proportion of negative reviews, with 10.72% and 14.70%, respectively, highlighting areas of potential improvement.



```
+--------------------+---------+------+-------------+------------------+
|    product_category|sentiment| count|total_reviews|        percentage|
+--------------------+---------+------+-------------+------------------+
|  Home Entertainment|  neutral|  3615|        36522|  9.89814358468868|
|               Tools| positive|  6302|         7510| 83.91478029294275|
|           Video DVD|  neutral| 72836|      1096788| 6.640845815235032|
|                Toys| positive| 49618|        57767| 85.89333010196133|
|         Video Games| positive| 12185|        15473| 78.75008078588509|
|               Video| positive| 37407|        46715| 80.07492240179815|
|               Video|  neutral|  3603|        46715| 7.712726105105426|
|         Video Games| negative|  2275|        15473|14.703031086408583|
|               Books| negative| 89937|       838729| 10.72301065063924|
|               Books| positive|685089|       838729| 81.68180663837784|
|               Music|  neutral| 50739|       778643| 6.516336755098292|
|           Video DVD| positive|916017|      1096788| 83.51814571275396|
|               Music| negative| 63506|       778643| 8.155984193012717|
|Digital_Music_Pur...| positive| 97014|       107855| 89.94854202401372|
|               Tools| negative|   599|         7510| 7.976031957390147|
|     Office Products| positive|  1742|         2313| 75.31344574146131|
|               Books|  neutral| 63703|       838729| 7.595182710982929|
|           Video DVD| negative|107935|      1096788|    9.841008472011|
|            Software| positive|    44|           53| 83.01886792452831|
|         Video Games|  neutral|  1013|        15473| 6.546888127706326|
+--------------------+---------+------+-------------+------------------+
only showing top 20 rows
```



Further Analysis

The results indicate strong sentiment trends, with certain categories like "Digital Music Purchases" and "Toys" displaying predominantly positive feedback. Categories with notable negative or neutral sentiment percentages may benefit from further investigation to identify underlying issues and improve customer satisfaction. This approach can help businesses prioritize efforts to enhance product performance and user experience.