Amazon Cell Phones Market Analysis – Exploratory Data Analysis and Machine Learning

1. Introduction

The cell phone marketplace on Amazon is vast and highly competitive, featuring thousands of listings from top global brands. With consumers increasingly relying on ratings, reviews, and pricing to make informed purchasing decisions, understanding the patterns in such data becomes essential for both sellers and platform strategists. This report presents a comprehensive data-driven analysis of Amazon's cell phone market. The aim is to identify the factors that most influence customer choices, segment products meaningfully, and provide actionable insights for optimization of listings, pricing strategies, and customer engagement.

2. Objectives

The primary goals of this project are:

- To explore the relationships between product attributes such as price, brand, rating, and customer feedback.
- To identify key drivers of consumer purchase decisions through regression analysis.
- To segment phones into logical clusters based on features like RAM, storage, and pricing using clustering algorithms.

- To understand how consumer sentiment, inferred from ratings, aligns with product pricing and performance.
- To derive actionable insights for improving product listings and customer targeting strategies.

3. Tools and Technologies Used

The project was implemented using Python in a Jupyter Notebook environment. The following libraries and frameworks were utilized:

- **Pandas** and **NumPy** for data manipulation and preparation.
- **Matplotlib** and **Seaborn** for data visualization.
- **Scikit-learn** for machine learning models including linear regression and clustering.
- **WordCloud** for visualizing common terms in reviews (optional).
- Additional libraries for warnings suppression and styling.

4. Dataset Description

The dataset used for this analysis was obtained from Kaggle and contains scraped data from Amazon's cell phone listings. It comprises:

- **Total records**: 3,351
- **Columns**: 19, including product name, brand, price, discount percentage, rating, number of ratings, operating system, RAM (GB), storage (GB), screen size, CPU model, and cellular technology.

Initial inspection revealed over 22% missing values, particularly in columns like `price_before_discount`, `CPU`, and `available_colors`, necessitating a structured data cleaning approach.

5. Data Preprocessing

Data preprocessing involved several steps:

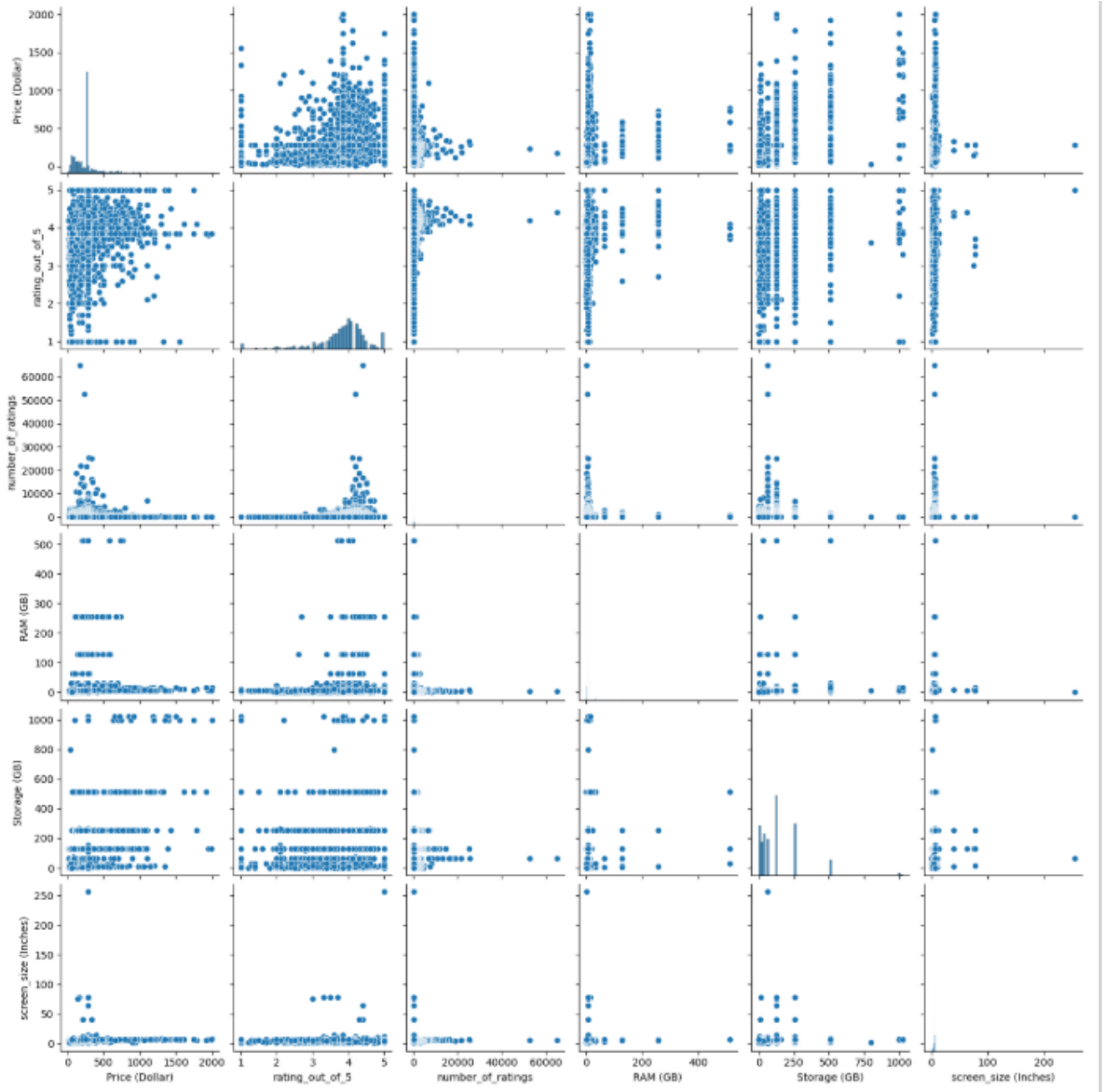
- **Numerical Imputation**: Missing numerical values (e.g., Price, RAM, Rating) were filled using the mean or median.
- **Categorical Imputation**: Categorical fields were filled with the mode or set to "Unknown".
- **Data Type Conversion**: Ensured consistency in column types for numeric and categorical processing.
- **Missing Data Analysis**: Visualized using heatmaps to identify sparsity patterns and evaluate data quality.

This step helped ensure that the data was clean, consistent, and ready for analysis.

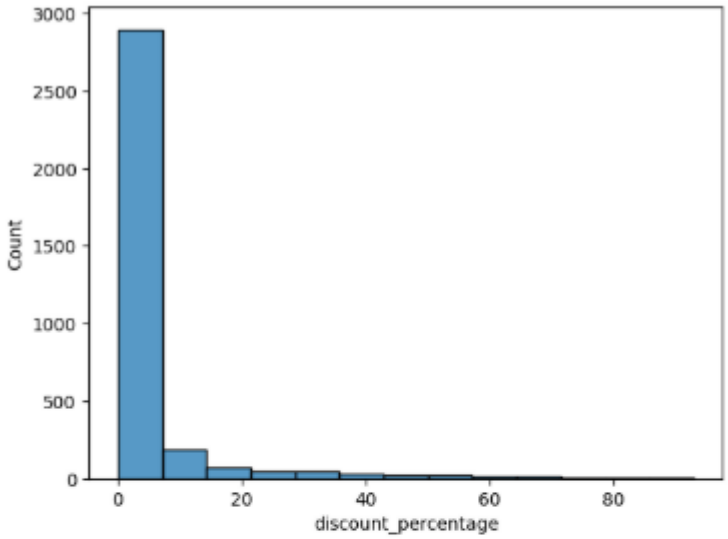
6. Exploratory Data Analysis (EDA)

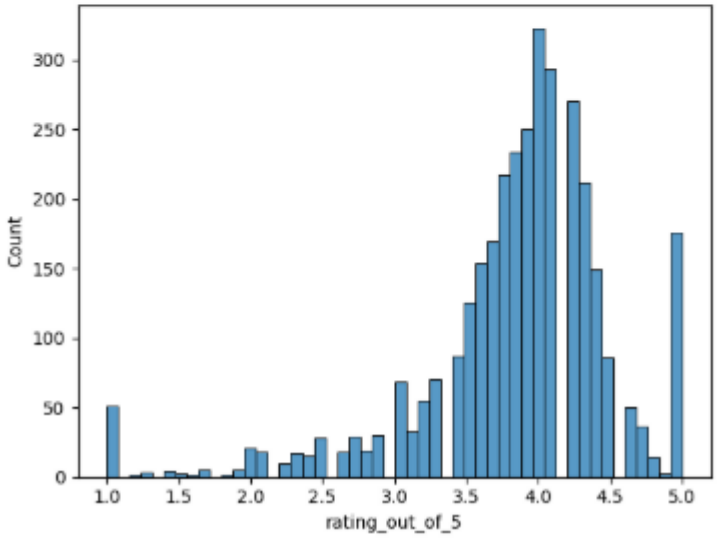
We performed extensive visual and statistical analysis to understand the dataset's structure and feature relationships:

- **Pairplot** of numeric variables revealed linear trends between price, rating, and storage.

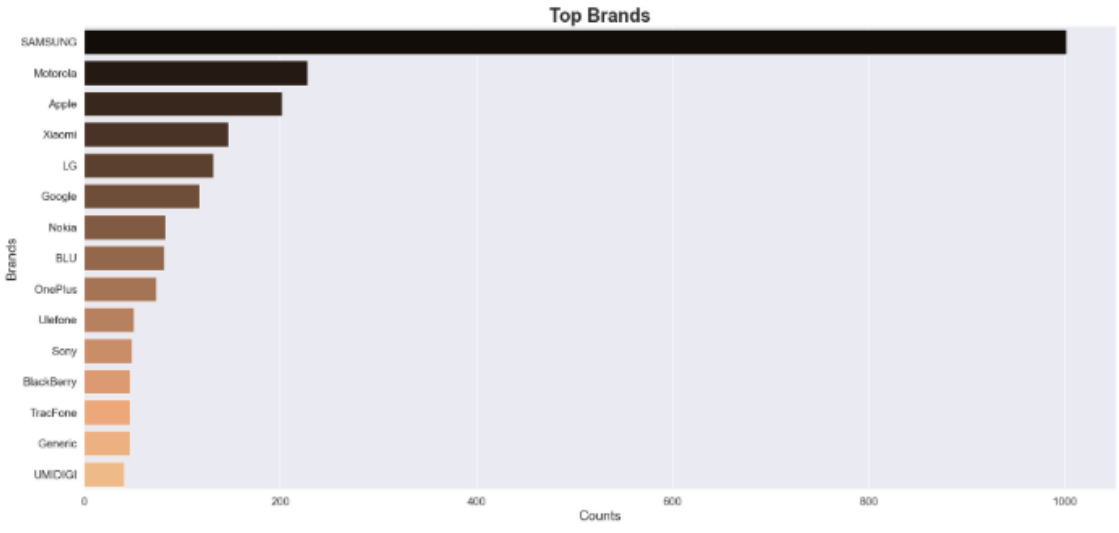


- **Histograms** showed the skewed nature of price and rating distributions.

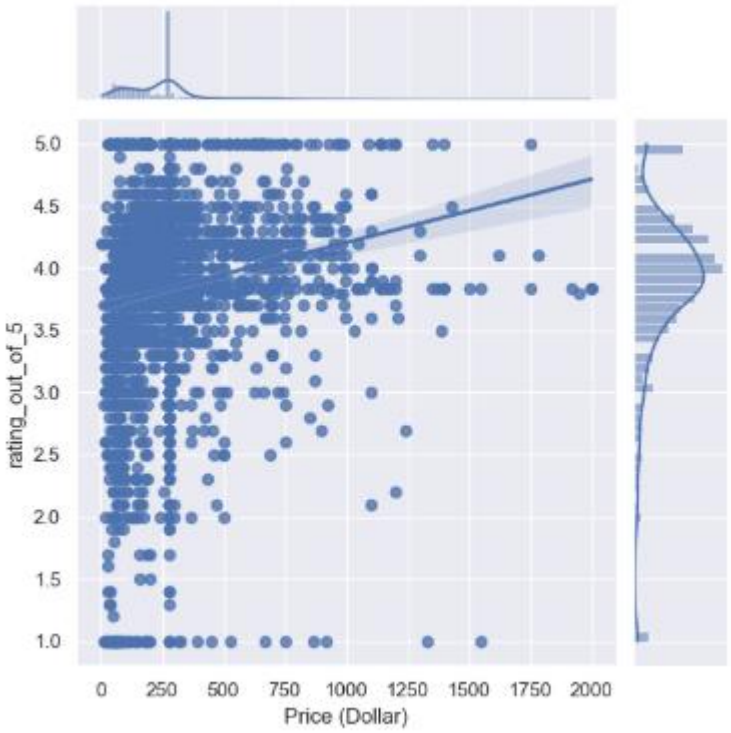




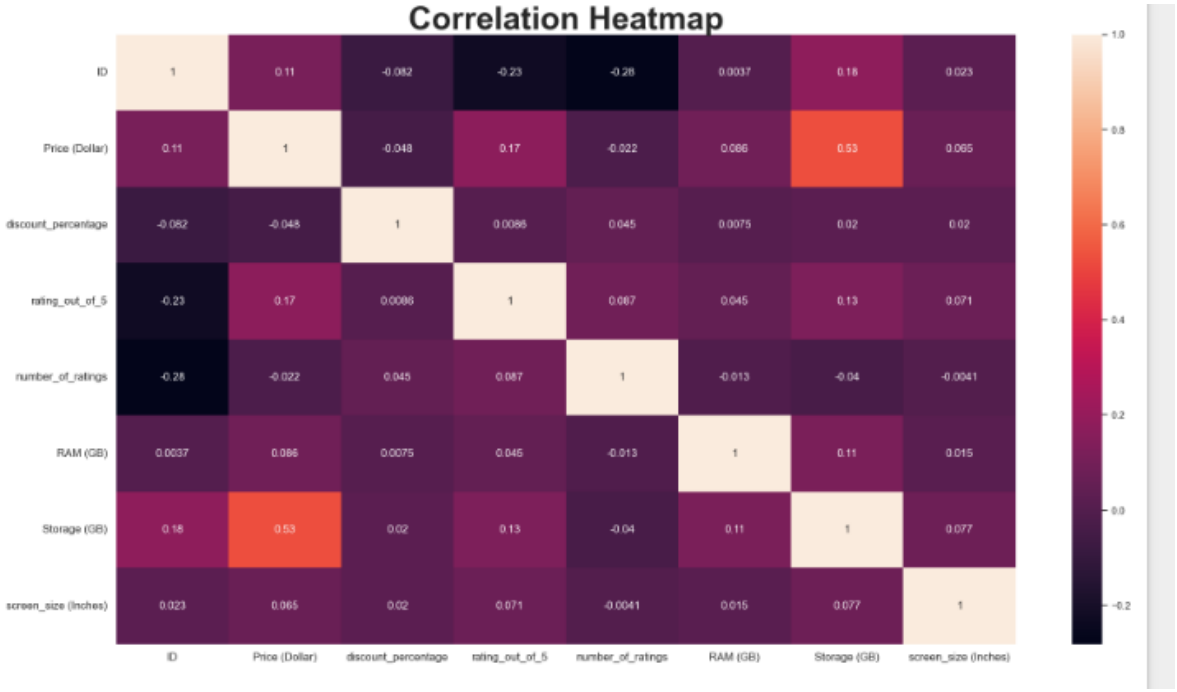
- **Bar charts** identified the top brands by count and price, highlighting Apple and Samsung as market leaders.



- **Line and scatter plots** examined pricing trends across product IDs and the correlation between price and ratings.



- **Correlation Heatmap** visually presented the relationships among numeric features.



These visual insights set the foundation for more complex modeling.

7. Regression Analysis

A multiple linear regression model was built to predict the price of a cell phone using:

- Rating (out of 5)
- Number of Ratings
- RAM (GB)
- Storage (GB)
- Discount Percentage

Model performance:

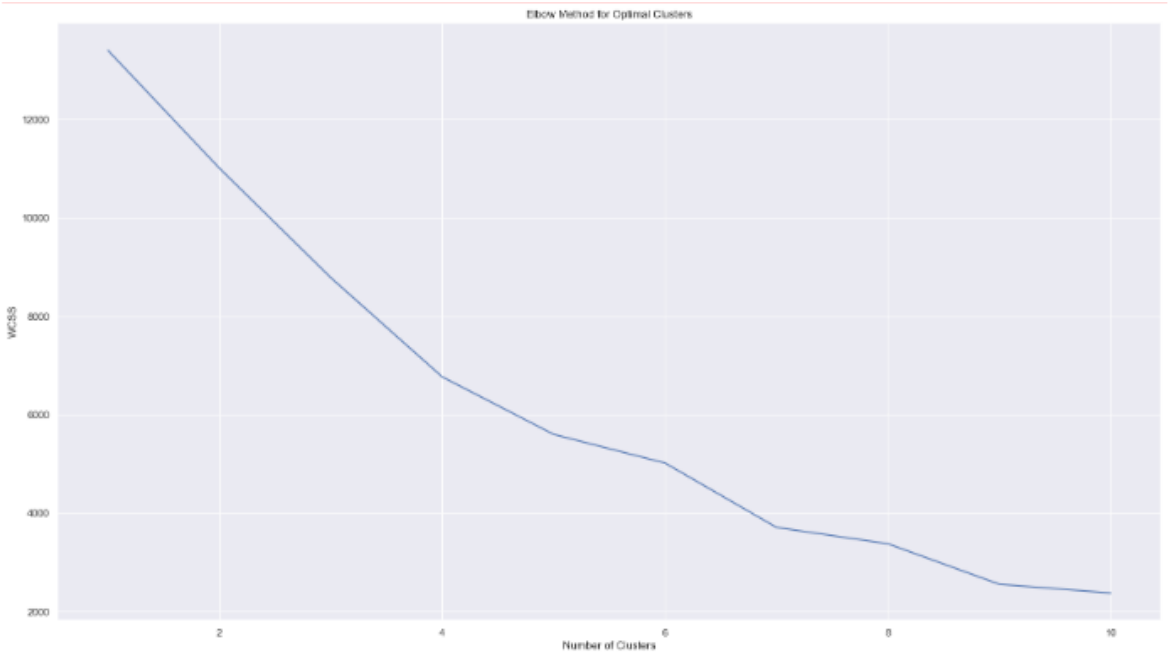
- **R-squared**: 0.305 (approximately 30% variance explained)
- **Key contributors**: Storage, RAM, and discount percentage were among the top features influencing price.

This model provides a useful baseline for estimating product price from attributes.

8. Clustering Analysis

We applied KMeans clustering on standardized features:

- Price
- RAM
- Rating
- Number of Ratings



Using the **Elbow Method**, the optimal number of clusters was determined to be 3. The clusters correspond to:

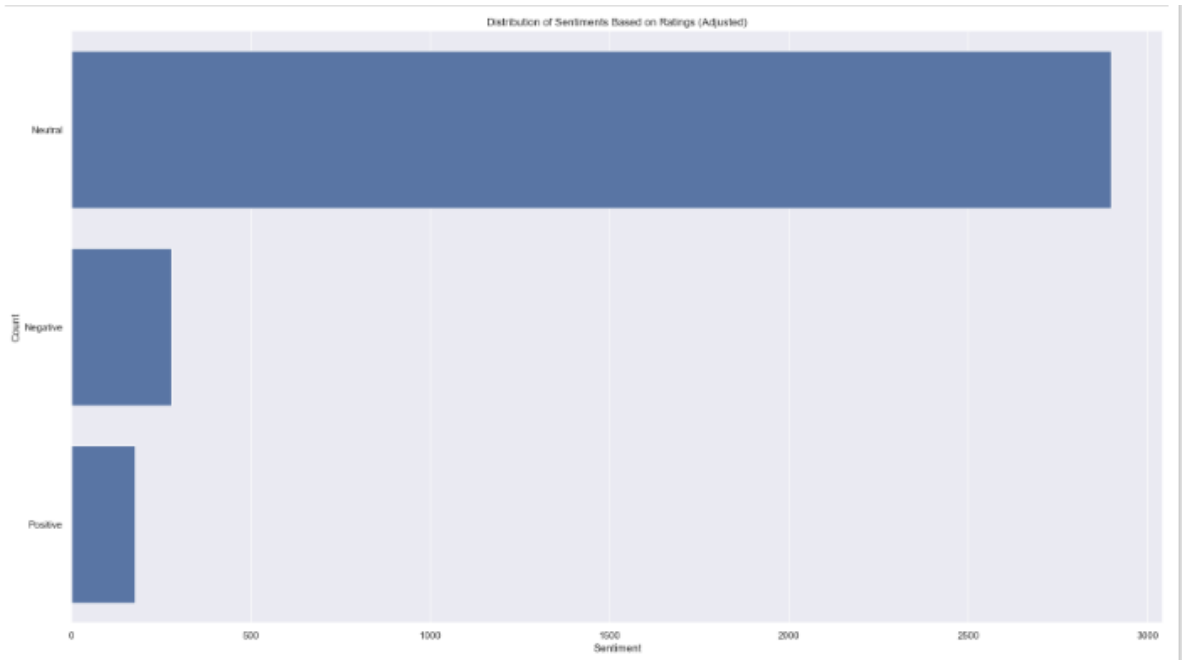
- **Budget Phones**
- **Mid-Range Phones**
- **Premium Phones**

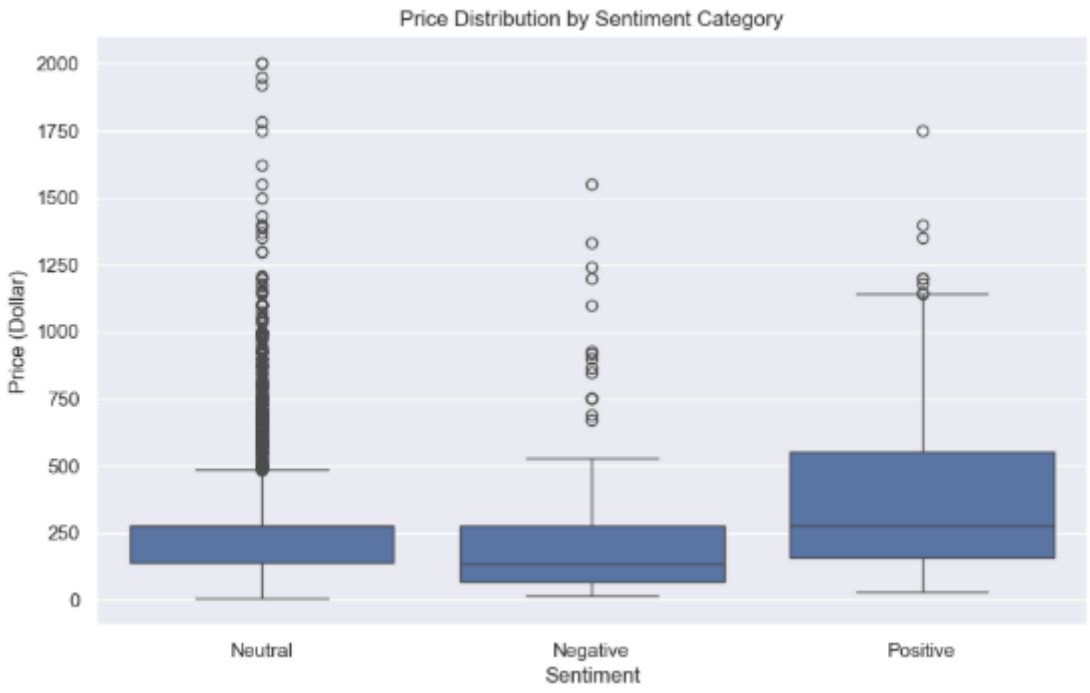
This segmentation can help sellers target the right audience with product-specific marketing strategies.

9. Sentiment Analysis

Since detailed textual reviews were not available, we inferred sentiment based on star ratings:

- **Positive**: 5 stars
- **Neutral**: 3 to 4.9 stars
- **Negative**: Below 3 stars





We observed that higher-priced phones were more likely to receive positive ratings. Sentiment-based analysis highlighted the relationship between product satisfaction and pricing tiers.

10. Key Insights

- **Brand Dominance**: Samsung and Apple dominate the listings both in volume and price.
- **Feature-Driven Pricing**: Storage capacity and RAM are the strongest predictors of price.
- **Discount Irrelevance**: Discounts had negligible correlation with better ratings or higher sales.
- **Clear Segmentation**: Clustering analysis effectively grouped products into logical pricing categories.

- **Rating Sentiment**: Positive ratings are commonly associated with higher-priced and higher-quality products.

11. Business Value and Applications

This analysis delivers high utility for:

- **E-commerce Sellers**: To optimize product pricing, highlight strong features, and plan targeted promotions.
- **Product Strategists**: To analyze feature demand and market segmentation.
- **Marketers**: To tailor campaigns for different product tiers identified via clustering.
- **Amazon or other Platforms**: To refine recommendation systems and enhance user experience with more personalized results.

12. Conclusion

This study provides a comprehensive understanding of the Amazon cell phone market through the lens of data analysis and machine learning. By exploring relationships between product features, predicting price, clustering similar items, and analyzing sentiment patterns, we derive actionable insights that can directly inform sales strategy, pricing optimization, and customer targeting. The combination of EDA, modeling, and visualization in this project demonstrates the power of data science to influence real-world decision-making in e-commerce.