Amazon India Headphones Market Data Analysis

Ishita Dingare

June 2025

Executive Summary

This project delivers a comprehensive analysis of the Amazon India headphones market using real-world data. Through a combination of SQL, Python and Power BI, I cleaned and structured the dataset, performed advanced statistical and machine learning analysis, and built interactive dashboards for business users. My findings reveal a highly competitive market with clear segmentation, strong brand dynamics, and actionable opportunities for both volume and premium strategies. The Power BI dashboard enables business stakeholders to explore these insights, supporting data-driven decisions in product positioning, pricing and brand strategy.

1. Business Problem and Objective

1.1 Objective:

To analyze the Amazon headphones market, uncover brand and product performance trends, understand pricing and value dynamics, and provide actionable recommendations for business growth.

1.2 Key Business Questions:

- Which brands and products dominate the market?
- How do pricing and customer ratings vary across segments?
- Where are the opportunities for value, premium, or volume strategies?
- How can data-driven insights inform product, pricing, and marketing decisions?

2. Data Collection and Preparation

2.1 Source:

Data was web-scraped from Amazon India, capturing product details, pricing, ratings, reviews, and brand information for headphones.

2.2 Cleaning:

- SQL and Python were used to standardize brand names, remove duplicates, fix web scraping errors, and handle the missing values.
- Final dataset: 723 products, 72 unique brands, with key columns: Title, Brand, Price, MRP, Discount, Rating, Review Count, Prime, ASIN, URL.

2.3 Feature Engineering:

- Added columns for Saving, Price Segment (Budget, Mid-range, Premium, Luxury), Value Score (rating per ₹1000), and product clusters.

3. Analytical Approach

3.1 SQL:

Initial data cleaning, deduplication, and summary statistics.

3.2 Python:

Advanced EDA, statistical correlation, clustering (KMeans), regression (Random Forest), and value analysis.

3.3 Power BI:

Interactive dashboards for executive overview, brand/product performance, price segmentation, and advanced insights.

4. Key Findings & Insights

4.1 Market Overview:

- **Market Size**: 723 unique products, 72 brands.
- **Price Range**: ₹99 - ₹29,990 (Avg: ₹1,892).
- **Average Rating**: 3.81/5.
- **Top Brands by Product Count**: boAt (159), Noise (60), Boult (60), ZEBRONICS (47), JBL (38).
- **Market Share**: boAt leads with 22% of products; top 10 brands control 70% of the market.

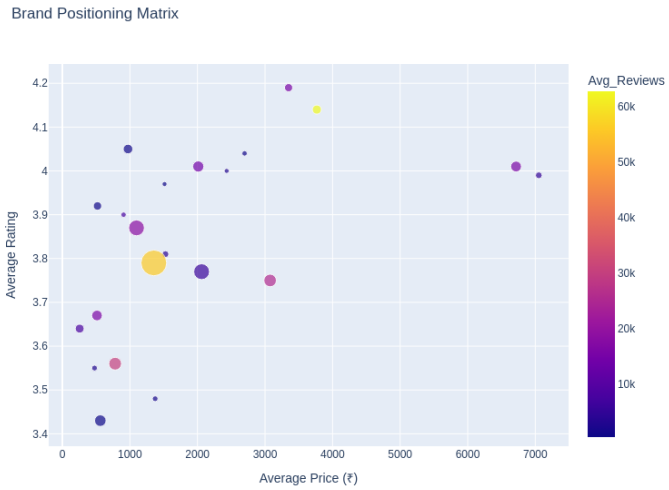

Figure 1: Overview

4.2 Brand & Product Performance:

- **Volume Leaders**: boAt dominates by product count and reviews engagement.
- **Premium Brands**: OnePlus (₹3,767, 4.14 rating), Sony (₹6,716, 4.01), soundcore (₹3,349, 4.19).
- **Value Brands**: Amazon Basics (₹562, 3.43), pTron (₹514, 3.67), Ambrane (₹256, 3.64).

- **Customer Engagement**: boAt and Noise have the highest review volumes, indicating strong brand presence.

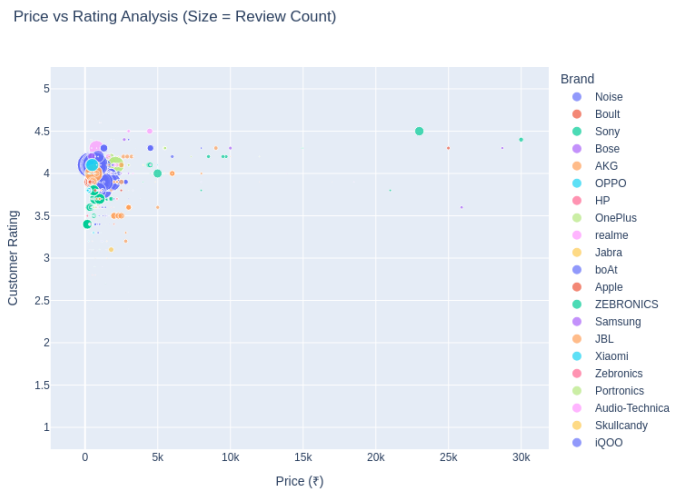

Figure 2: Price vs. Rating

4.3 Price Segmentation & Value:

- **Product Distribution**:
  - Mid-range (₹500-₹1,500): 345 products (48% of the market).
  - Premium (₹1,500-₹5,000): 197 products (27%).
  - Budget (<₹500): 138 products (19%).
  - Luxury (>₹5,000): 43 products (6%).
- **Quality-Price Correlation**:
  - Average rating increases with price: Budget (3.69), Mid-range (3.75), Premium (3.94), Luxury (4.10).
  - Value Score drops as price increases: Budget (11.36), Luxury (0.47).
- **Customer Engagement:**
  - Budget and mid-range products have the highest average review counts, luxury products have lower engagement but higher satisfaction.

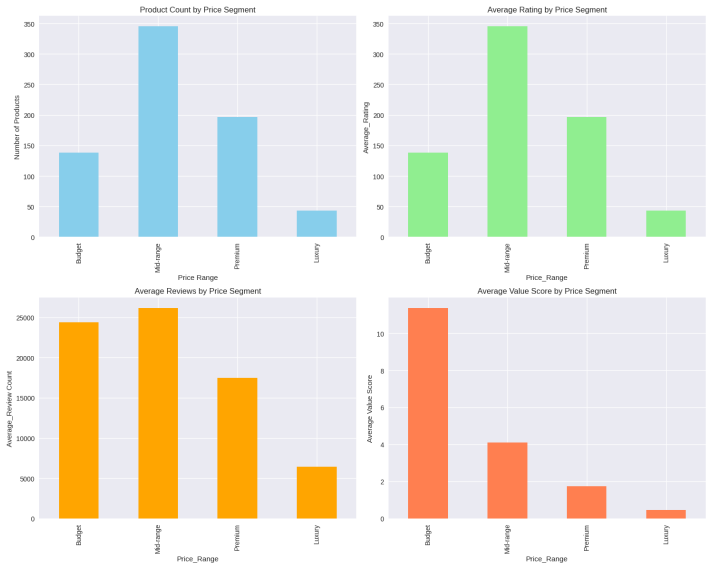

Figure 3: Market Segmentation Analysis

4.4 Advanced Insights:

- **Correlation Analysis**:
  - Moderate positive correlation between price and rating (r=0.22, p¡0.001).
  - Weak negative correlation between price and review count (r = -0.07).
  - Value Score is strongly negatively correlated with price (r = -0.39): Cheaper products offer better rating-per-rupee.

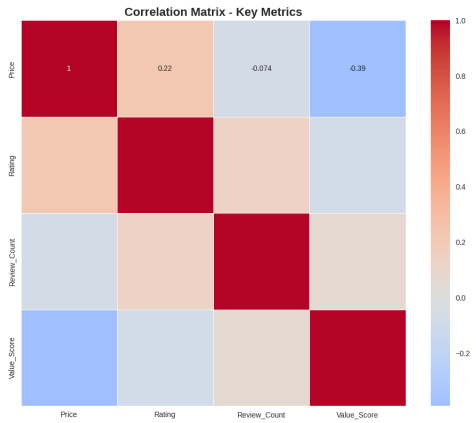

Figure 4: Correlation Analysis

- **Clustering**: Four clusters identified:
  - Cluster 0: Mid-priced, high rating, very high reviews (popular, good value).
  - Cluster 1: Upper mid-range, moderate rating/reviews.
  - Cluster 2: Premium/luxury, highest price/rating, low value score.
  - Cluster 3: Budget, moderate rating, high reviews, highest value score.

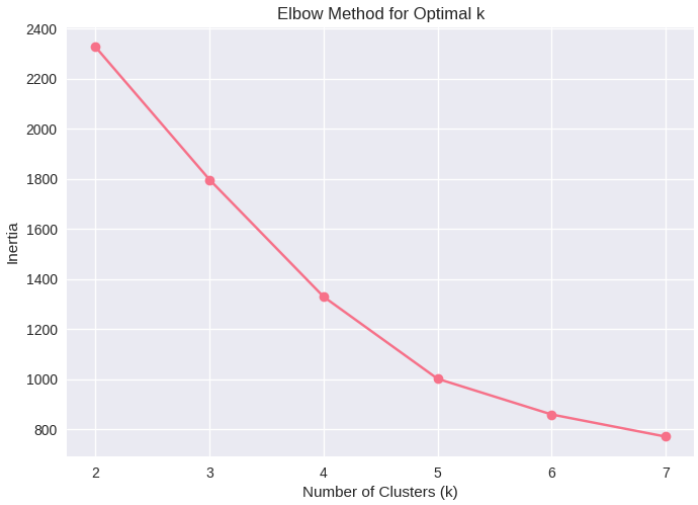

Figure 5: Elbow Method Plot

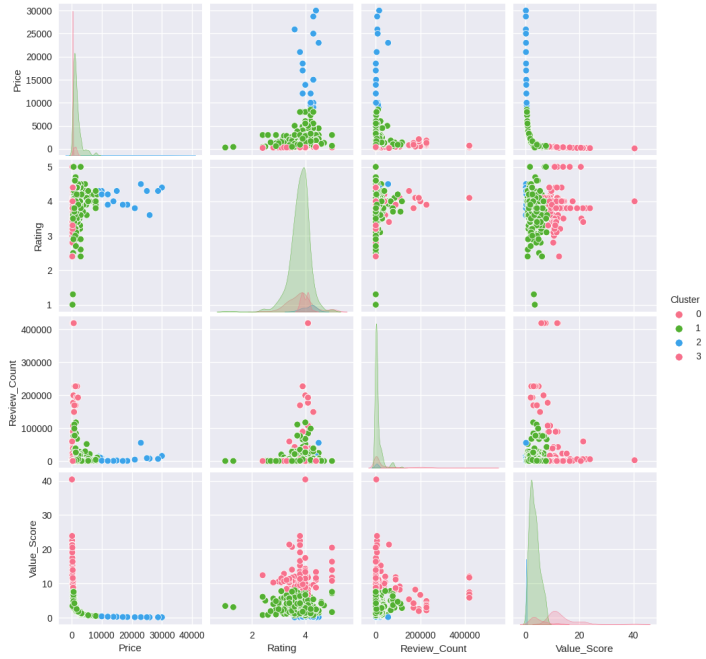

Figure 6: Cluster Analysis

- **Brand vs. Price Segment**:
  - Premium brands (OnePlus, Sony, soundcore) lead in ratings and price.
  - Value brands (truke, Portonics) deliver high ratings at low prices.
- **Prime vs. Non-Prime**:
  - Prime products (89% of market) command higher prices and ratings.

4.5 Price Prediction Modeling:

Random Forest Regression:

- Test R^2 up to 0.51 with brand, rating, and review count as features.

- Rating and brand are the most important predictors of price, but much price variation remains unexplained due to product features not captured in the dataset.

5. Recommendations

5.1 For New Entrants:

- **Target the mid-range segment (₹500 - ₹1,500)**: Highest product count and customer engagement.
- **Focus on achieving 4.0+ ratings** to compete in premium segments.
- **Consider budget entry for volume strategies**—high review counts indicate strong demand.

5.2 For Existing Brands:

- **Premium migration:** Brands in the mid-range should improve quality to move into premium segments.
- **Exploit quality gaps:** Bridge the rating gap between segments and differentiation.
- **Leverage customer reviews:** Use feedback to improve products and boost ratings.

5.3 For Pricing Strategy:

- **₹1,000 - ₹1,500:** Highly competitive, requires strong differentiation.
- **₹2,000 - ₹3,000:** Premium positioning quality focus.
- **Under ₹500:** Volume play; maintain acceptable quality standards.

5.4 For Value Optimization:

- **Ultra-budget products (<₹200:)** Can achieve exceptional value scores and mass-market appeal.
- **Premium products:** Must justify higher prices with superior ratings and features.

6. Dashboard Support

6.1 Power BI dashboards allow business users to:

- Explore market share and brand performance interactively.
- Analyze product distribution, ratings, reviews, and value by price segment.
- Identify top-performing products and clusters.
- Drill down by brand, price segment, or Prime status for targeted insights.

6.2 Visuals include:

KPI cards, bar/column charts, scatter plots, heatmaps, and cluster analysis, all designed for executive decision-making.

7. Limitations & Next Steps

7.1 Data limitations:

- Web scraping errors and missing values were addressed, but some product features (e.g., technical specs, launch date) are not captured.
- Price prediction is limited by available features; including more product attributes could improve accuracy.

7.2 Future work:

- Add time-series analysis for price and rating trends.
- Integrate sales volume or revenue data for deeper business insights.
- Apply NLP on product reviews for sentiment analysis.

8. Appendix

8.1 Dashboard screenshots:

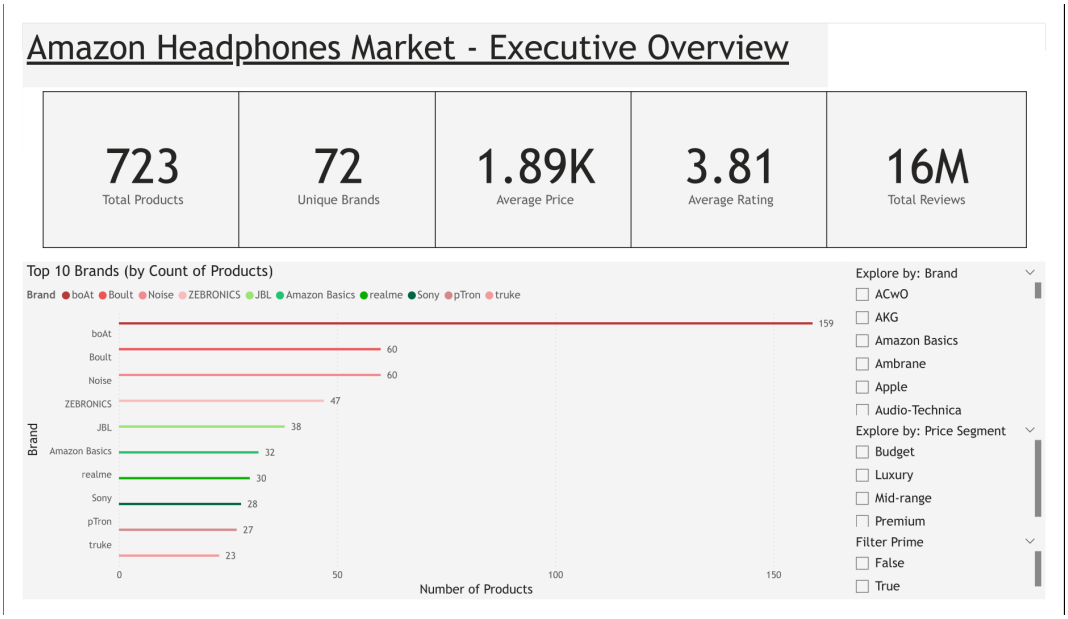

Figure 7: Executive Overview

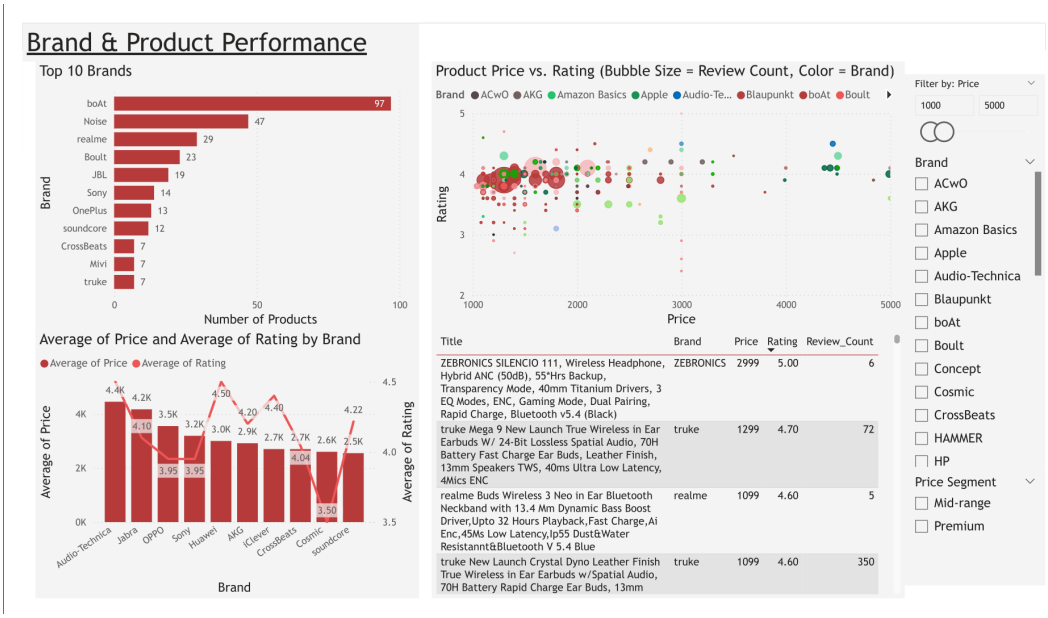

Figure 8: Brand & Product Performance

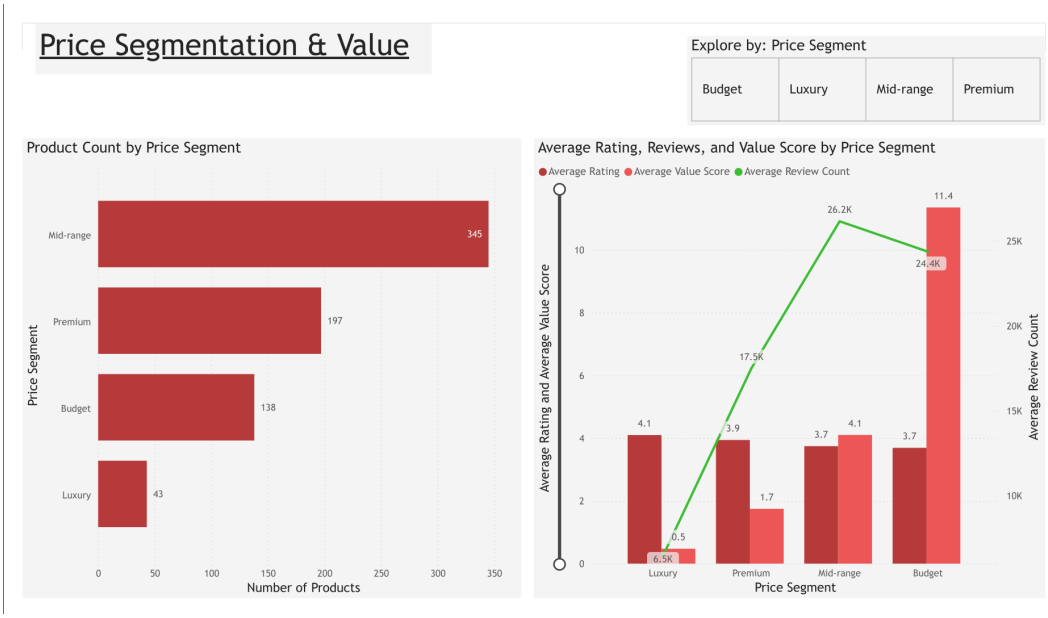

Figure 9: Price Segmentation & Value

Figure 10: Advanced Insights

8.2 Sample codes:

1. Sample SQL code for Data Cleaning

```sql
-- Standardizing brand names
UPDATE amazon_data
SET "Brand" = CASE
    WHEN LOWER("Brand") = 'boat' THEN 'boAt'
    WHEN LOWER("Brand") = 'oneplus' THEN 'OnePlus'
    WHEN LOWER("Brand") = 'amazon' THEN 'Amazon Basics'
    WHEN LOWER("Brand") = 'zebronics' THEN 'ZEBRONICS'
    WHEN "Brand" IN ('Q', 'W20', 'Mustang') THEN 'Boult'
    ELSE "Brand"
END;

-- Deduplication using ASIN (unique for each product)
CREATE TABLE cleaned_amazon_data AS
SELECT
    "Title", "Brand", "Price", "MRP", "Discount",
    "Rating", "Review_Count", "Prime", "ASIN", "URL"
FROM (
    SELECT *, ROW_NUMBER() OVER (PARTITION BY "ASIN" ORDER BY "Title") as rn
    FROM amazon_data
    WHERE "ASIN" IS NOT NULL AND "ASIN" != ''
) ranked
WHERE rn = 1;

-- Clean Price column
UPDATE cleaned_amazon_data
SET "Price" = REGEXP_REPLACE("Price", '[^0-9]', '', 'g')::INTEGER
WHERE "Price" IS NOT NULL AND "Price" ~ '[0-9]';

-- Clean Rating column
UPDATE cleaned_amazon_data
SET "Rating" = SPLIT_PART("Rating", ' ', 1)::FLOAT
WHERE "Rating" IS NOT NULL AND "Rating" ~ '^[0-9\.]+';
```

2. Sample Python code for Feature Engineering and Analysis

```python
# Feature engineering
df['Price_Range'] = pd.cut(df['Price'],
    bins=[0, 500, 1500, 5000, float('inf')],
    labels=['Budget', 'Mid-range', 'Premium', 'Luxury']
)
df['Value_Score'] = (df['Rating'] / (df['Price'] / 1000)).round(3)

# Correlation analysis
numeric_cols = ['Price', 'Rating', 'Review_Count', 'Value_Score']
correlation_matrix = df[numeric_cols].corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', center=0)

# Random Forest regression
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer

preprocessor = ColumnTransformer([
    ('brand', OneHotEncoder(handle_unknown='ignore'), ['Brand'])
], remainder='passthrough')

X = df[['Brand', 'Rating', 'Review_Count']]
y = df['Price']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state
    =42)

X_train_processed = preprocessor.fit_transform(X_train)
X_test_processed = preprocessor.transform(X_test)

rf = RandomForestRegressor(n_estimators=100, max_depth=20, min_samples_leaf=2,
    max_features=0.8, random_state=42)
rf.fit(X_train_processed, y_train)
```

8.3 Data dictionary:

Column	Description
Title	Product title/name as listed on Amazon
Brand	Brand name (standardized, e.g., "boAt", "OnePlus")
Price	Current selling price (in ₹)
MRP	Maximum Retail Price (in ₹)
Discount	Discount percentage from MRP (integer)
Rating	Customer rating (out of 5, float)
Review_Count	Number of customer reviews (integer)
Prime	Whether the product is Amazon Prime eligible (TRUE/FALSE)
ASIN	Amazon Standard Identification Number (unique product identifier)
URL	Product page URL on Amazon
Savings	Amount saved from MRP (MRP - Price, in ₹)
Price_Segment	Price category: Budget, Mid-range, Premium, or Luxury (engineered feature)
Value_Score	Rating per ₹1000 spent (engineered feature: Rating / (Price/1000))

9. Conclusion

This project demonstrates a full-cycle, real-world data analytics workflow:

- Data collection, cleaning, and feature engineering
- Advanced analysis (statistical, clustering, regression)
- Business-focused visualization and reporting

The insights and dashboards produced here enable better strategic decisions in product positioning, pricing, and brand management for the Amazon headphones market.