Market Traction Segmentation and Intervention Prioritization in True Wireless Earphones

An Unsupervised Learning Analysis Based on Observable Market Signals

SOHINI MANDAL

Abstract

E-commerce platforms host large product catalogs in which observable market behavior varies across price levels, user ratings, and cumulative engagement. When direct information on sales, exposure, or product lifecycle stage is unavailable, publicly observable signals can be used to examine relative market traction. This study applies unsupervised learning to segment true wireless earphones listed on Flipkart using price variables, cumulative review volume, and user ratings.

After data cleaning, transformation, and standardization, hierarchical clustering and K-Means segmentation were performed, with the number of clusters validated using structural and inertia-based criteria. Pairwise feature-space analysis, principal component analysis (PCA), and cluster-wise summary statistics indicate a continuous and overlapping market structure. Within this structure, products display distinct configurations of price, engagement, and ratings associated with relatively weaker observable market traction. The resulting clusters serve as relative monitoring and investigation priority groups based solely on observable market signals.

Introduction

Background

Product performance on e-commerce platforms is shaped by multiple interacting factors, including pricing, consumer engagement, and perceived quality. While platforms often possess rich behavioral data internally, analysts frequently rely on publicly observable signals such as ratings and review counts to infer market traction. These signals, though imperfect, provide valuable insight into relative product visibility and acceptance.

Motivation

Low observed market traction in public signals does not necessarily imply poor underlying product quality. Products may appear to underperform because of limited exposure, weak discoverability, or early lifecycle stages. Exploring whether products are more consistent with under-exposure or with structural weaknesses in their observable signals is therefore an important input to pricing, promotion, and catalog optimization.

Problem Statement

Given all products currently visible on the platform, which ones exhibit relatively weaker observable market traction signals and may require closer monitoring or intervention?

Objectives

The objectives of this study are:

- To segment true wireless earphones based on observable market signals such as price, ratings, and review volume.
- To identify product segments exhibiting relatively weaker observable market traction.
- To explore whether the signal patterns in these segments are more consistent with under-exposure or with structural weaknesses.
- To support evidence-based intervention prioritization using unsupervised learning techniques.

Dataset Description

Dataset name: Flipkart Earphones

Dataset link: https://www.kaggle.com/datasets/peshimaammuzammil/flipkart-earphones

About this dataset: This is a pre-crawled dataset, taken as subset of a bigger data set (more than 5.8 million products) that was created by extracting data from Flipkart.com, a leading Indian eCommerce store.

Variables present:

- **company:** The name of the company that makes the product.
- **name:** The name of the product.
- **color:** The color of the product.
- **type:** The type of product.
- **ratings:** The ratings of the product by users.
- **people_review:** The number of people who have reviewed the product.
- **offer_price:** The offer price of the product.
- **real_price:** The real price of the product.
- **offer:** The offer that is available for the product.

Data Cleaning and Preprocessing

Only True Wireless Earphones were retained for analysis to ensure product comparability and avoid confounding effects arising from heterogeneous product types.

The original dataset contained 414 observations and 9 variables, with several numeric fields stored as strings.

- Review Counts: *people_review was* cleaned by removing commas and converting to integers.
- Missing Values: Missing ratings were imputed with 0.
- Price Variables: *offer_price* and *real_price* were stripped of currency symbols and commas, then converted to float.

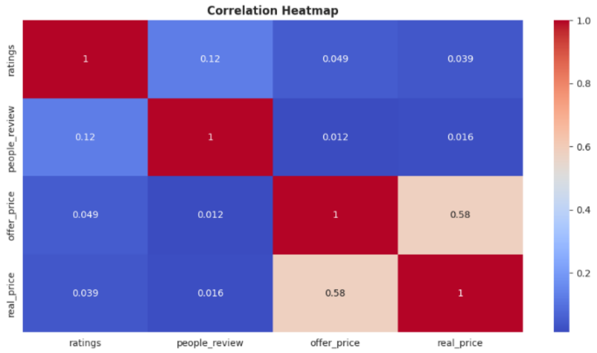
The cleaned dataset now contains 414 records and 7 variables(as *offer* column was dropped because it was redundant), 3 categorical (company, name, color) and 4 numerical (*ratings*, *people_review*, *offer_price, real_price*), with no missing values, ready for analysis.

Distributional Characteristics

Statistic	Ratings	People Review	Offer Price (₹)	Real Price (₹)
Count	414.00	414.00	414.00	414.00
Mean	3.88	64,185.45	946.53	3,640.47
Standard Deviation	0.29	204,284.00	952.78	2,235.18
Minimum	2.00	2.00	235.00	866.00
25th Percentile (Q1)	3.70	359.75	546.00	1,999.00
Median (Q2)	3.90	6,182.00	799.00	2,999.00
75th Percentile (Q3)	4.10	45,808.75	1,099.00	4,499.00
Maximum	5.00	1,345,614.00	16,999.00	26,300.00

- **Ratings** show low variability and cluster around higher values, indicating generally positive customer feedback.
- **Review counts** are extremely right-skewed, with a small number of products dominating visibility.
- **Offer and real prices** exhibit moderate to high dispersion with long upper tails, reflecting a mix of budget and premium products.

Correlation Structure



Correlation analysis reveals:

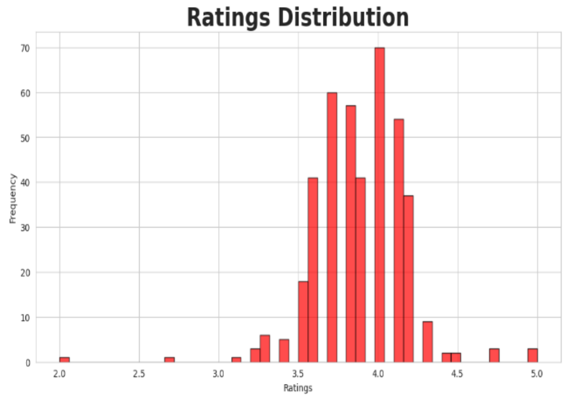
- Very weak relationships between ratings and both price and review count.
- Near-zero correlation between people review and price.
- A moderate positive correlation between offer price and real price.

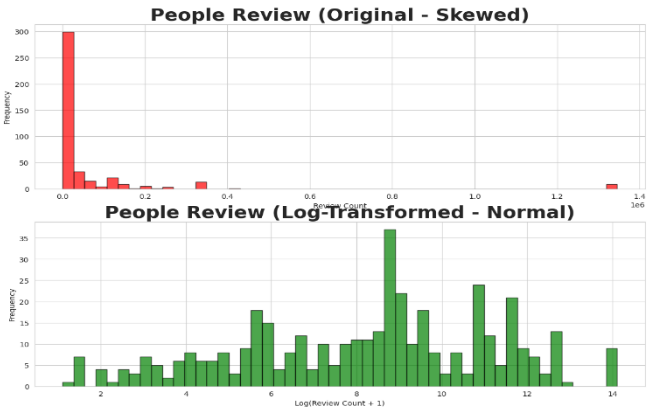
These results indicate that price, popularity, and perceived quality provide largely independent information.

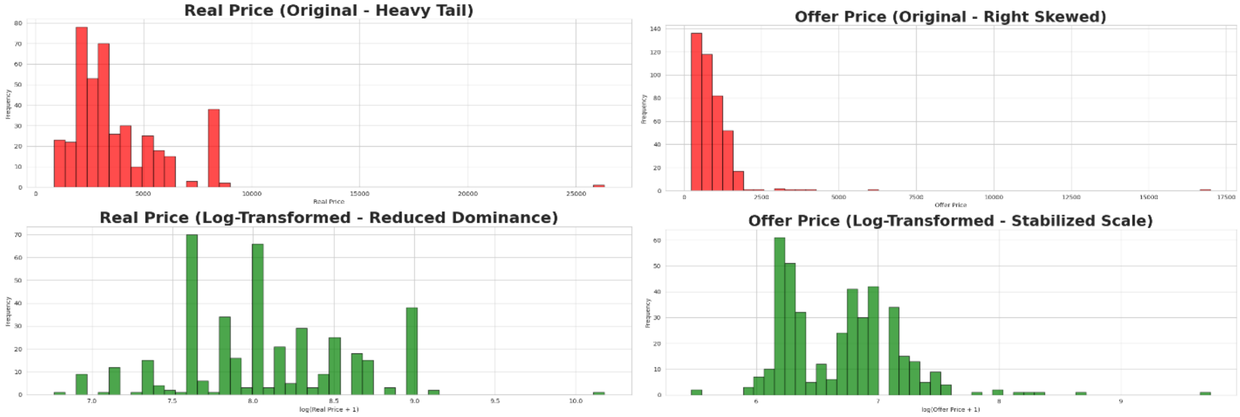
Feature Selection

We retained numeric features relevant to market traction: ratings, people_review, offer_price, and real_price. Categorical/text features (company, name, color, type) and the textual offer were excluded from clustering, as they are redundant given the price fields.

Feature Transformation and Scaling



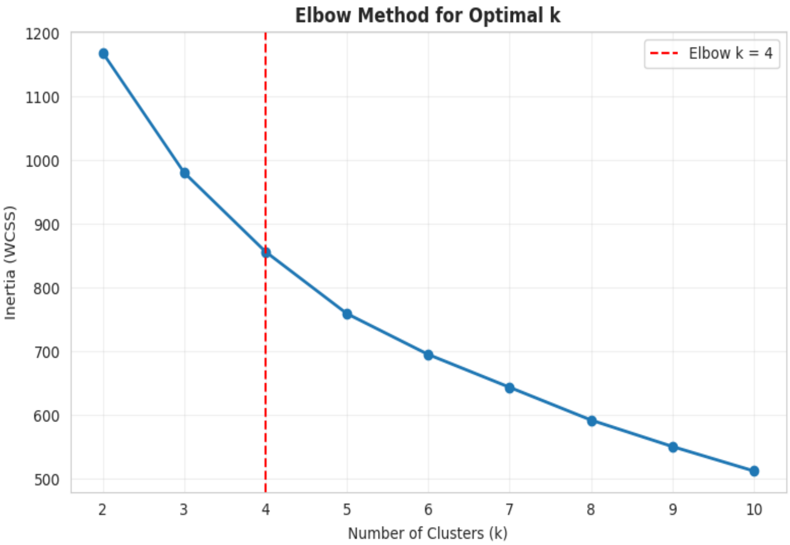




Strong skewness and extreme values motivated the application of logarithmic transformations to review counts and price variables. Log transformation stabilizes variance, reduces the dominance of extreme values, and enables meaningful segmentation without discarding observations.

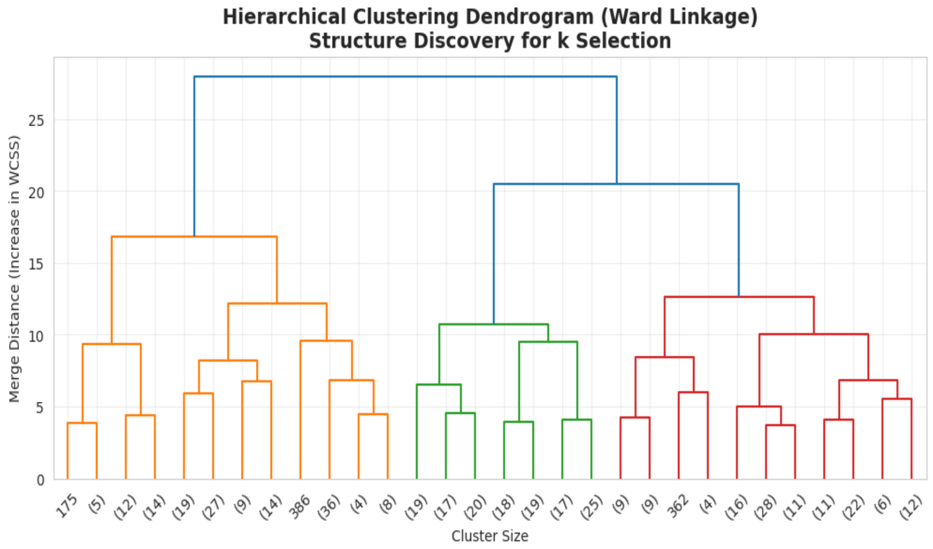
All clustering features were subsequently standardized to ensure equal contribution in distance-based algorithms.

Determining the Optimal Number of Clusters (Elbow Method)



We ran K-Means clustering (after scaling) for k = 2 to 10 and computed the within-cluster sum of squares (WCSS) for each. The WCSS plot showed a noticeable “elbow” at k = 4, indicating that adding more clusters beyond 4 yields diminishing returns in reducing WCSS. This suggests four clusters as an initial choice.

Hierarchical Clustering and Dendrogram



We applied agglomerative hierarchical clustering with Ward's linkage on the same standardized data. The resulting dendrogram (truncated to the last 30 merges for clarity) revealed a large jump in distance when merging beyond four clusters. This indicates that a four-cluster solution preserves more homogeneous groups, while merging further would combine dissimilar products. The hierarchical analysis thus validated k = 4 as a defensible segmentation, providing an evidence-based justification for choosing four clusters rather than an arbitrary number.

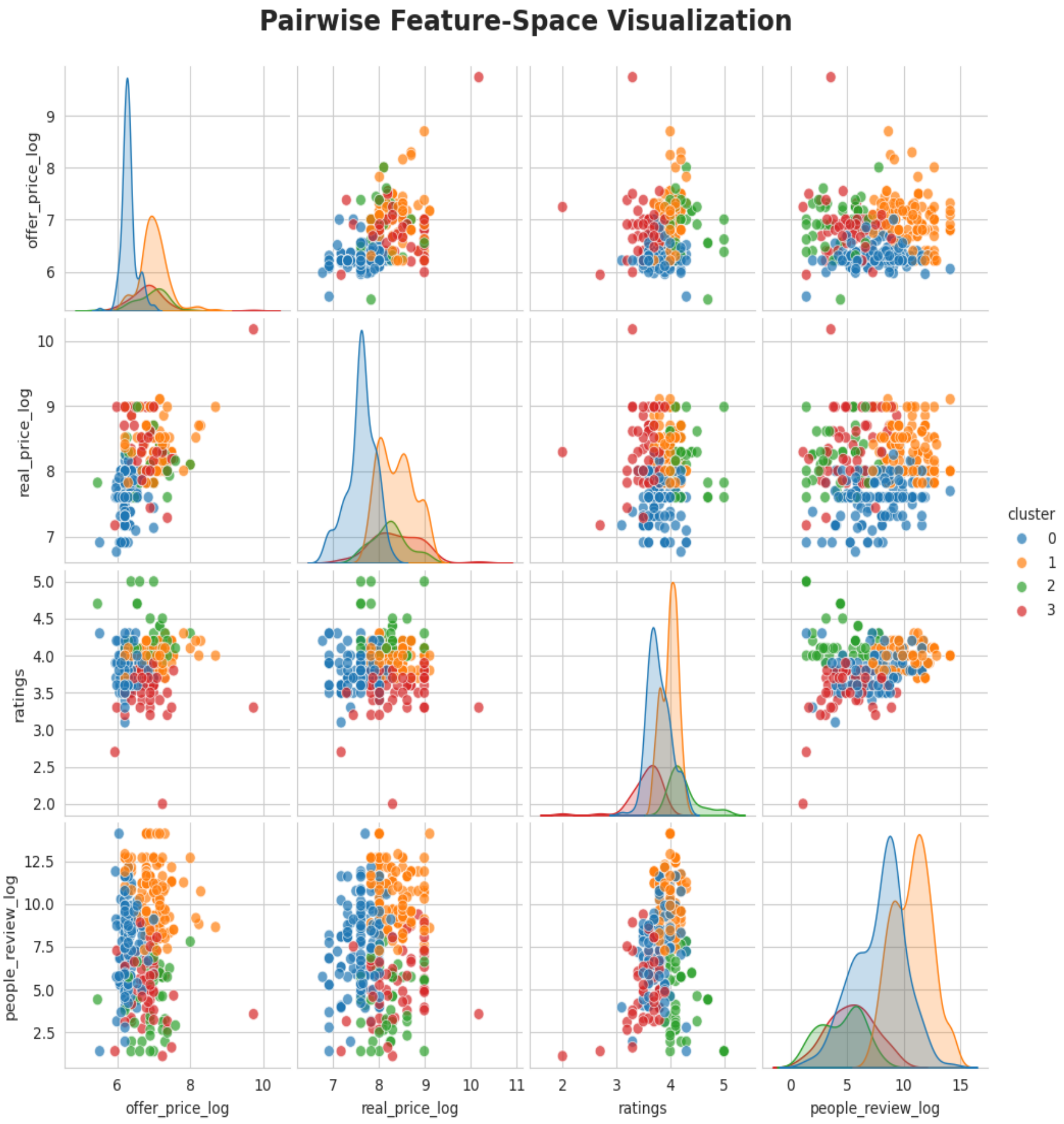
Linkage Computation using Ward Method: Ward linkage minimizes the variance within clusters during merging. It is especially effective for numeric, continuous variables, as it tends to produce compact, spherical clusters suitable for K-Means initialization.

K-Means Clustering and Final Segmentation

To identify natural groupings within the dataset, K-Means clustering was applied using the previously determined optimal number of clusters. Prior to clustering, the selected features were scaled to ensure that differences in magnitude did not disproportionately influence the distance calculations. K-Means is an iterative, centroid-based algorithm that partitions data into K distinct, non-overlapping clusters by minimizing the within-cluster variance (inertia). The algorithm operates in the following steps: initially, K centroids are chosen using smart initialization (k-means++) to improve convergence and reduce the likelihood of suboptimal clustering. Each data point is then assigned to the nearest centroid based on Euclidean distance, forming clusters. Centroids are subsequently recalculated as the mean of all points assigned to each cluster, and the assignment-update steps are repeated until the cluster assignments stabilize or changes in centroids become negligible.

After fitting the algorithm, each observation was assigned a cluster label, creating a final segmentation of the dataset. This segmentation facilitates meaningful analysis of patterns and similarities within clusters, enabling targeted insights for business or product-level decisions. K-Means is widely used due to its simplicity, computational efficiency, and effectiveness in revealing inherent structures in data.

Pairwise Feature-Space Analysis



We generated scatterplots of each variable pair, colored by cluster, to examine cluster overlap and separability. Key observations include:

- **Price vs. Price (Log-Offer vs. Log-Real): As** expected, there is a strong positive relationship: more expensive products tend to have higher offer prices. Cluster 0 points lie mostly in the lower-left (low prices), Cluster 1 shifts right (higher prices), and Clusters 2 and 3 occupy intermediate to high ranges. However, the scatter is broad at higher prices, reflecting varied discounting. Thus, price provides a *gradient* but does not strictly separate clusters.

- **Ratings vs. Price:** There is almost no clear trend between rating and price. High-priced products do not consistently earn higher ratings; ratings cluster tightly between ~3.5 and ~4.5 across all price levels. For example, both cheap and expensive products can have ratings around 4.0. Cluster 2 shows slightly higher ratings at mid-to-high prices, while Cluster 3 has notably lower ratings at higher prices. Overall, price alone is a poor predictor of quality in terms of user ratings.
- **Reviews vs. Price:** The number of reviews is broadly scattered across all prices. Cluster 1 tends to have the highest review counts (shifted upward in review volume), consistent with their market visibility. Cluster 0 has moderate reviews, while Clusters 2 and 3 have very low reviews across price levels. This indicates that *popularity* (as proxied by reviews) is only weakly related to price and varies cluster by cluster.
- **Ratings vs. Reviews:** Ratings and reviews show no strong association. We see highly-rated products with few reviews (Cluster 2) and highly-reviewed products with only moderate ratings (Cluster 1). All clusters overlap in this space, reinforcing that customer satisfaction and engagement are distinct dimensions.
- **Marginal Distributions:** The diagonal plots (density estimates) illustrate that even after log transformation, offer/real prices and review counts remain somewhat skewed but have overlapping distributions across clusters. No single variable shows a cluster-specific peak; instead clusters shift the location of similar distributional shapes.

Taken together, these analyses confirm that no single feature cleanly separates the clusters. Instead, clusters reflect multivariate patterns. The market resembles a continuum of products where cluster membership arises from combined levels of price, popularity, and quality. For instance, Cluster 1 (popular premium products) and Cluster 0 (budget staples) both include a range of reviews and ratings, but are distinguished by price levels. Clusters 2 and 3 are isolated more by their extreme values in reviews and ratings than by price alone.

Cluster-wise Statistical Summary (Mean and Variance)

Cluster-wise Mean Statistics

Cluster	Offer Price (₹)	Real Price (₹)	Ratings	People Review
0	561.10	2,090.71	3.79	21,314.82
1	1,198.47	4,663.08	3.98	152,478.96
2	1,131.48	4,046.35	4.24	312.22
3	1,247.93	5,113.20	3.55	1,106.96

Cluster-wise Variance Statistics

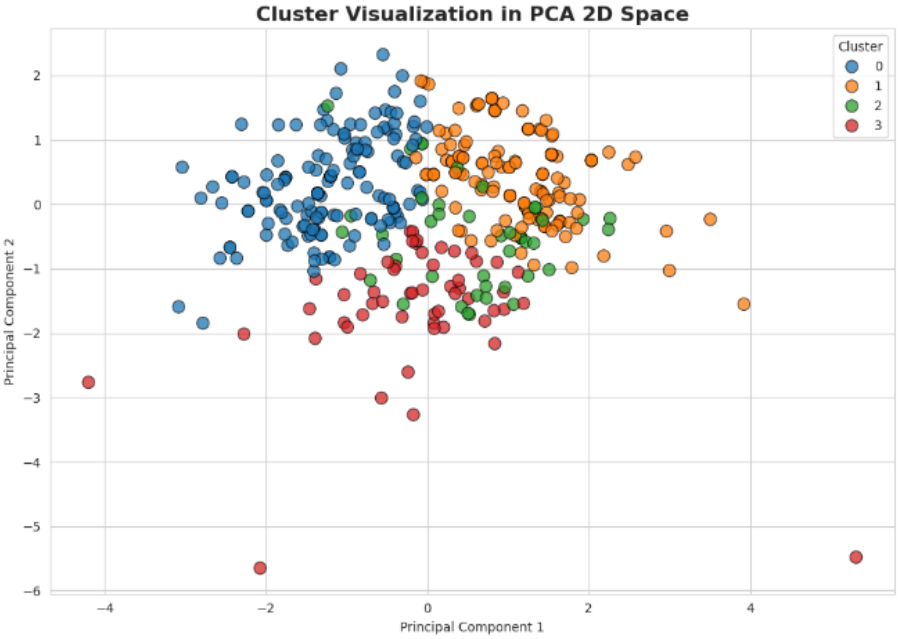
Cluster	Offer Price Variance	Real Price Variance	Ratings Variance	People Review Variance
0	15,164.64	362,217.60	0.0461	11,685,170,000
1	423,954.50	3,329,175.00	0.0217	89,751,720,000
2	222,511.50	2,909,105.00	0.0771	251,014.90
3	4,878,742.00	13,213,340.00	0.0961	5,087,947.00

Cluster-wise means and variances were computed to summarize average observable signal levels and internal dispersion within each cluster. The statistics describe central tendencies and heterogeneity conditional on cluster membership, without implying prevalence or dominance.

Across clusters, mean price levels differ substantially, while ratings vary within a relatively narrow range. In contrast, cumulative review counts exhibit large differences in both mean and variance, indicating that observable engagement varies more sharply than price or ratings. Variance estimates further show that some clusters display relatively homogeneous engagement levels, whereas others contain products with widely dispersed review counts and prices.

Overall, the mean–variance summaries complement the clustering and feature-space analyses by highlighting differences in average observable signals and internal heterogeneity, reinforcing the interpretation of the segmentation as soft and overlapping rather than sharply separated groups.

Principal Component Analysis (PCA)



To visualize the overall structure, we applied PCA to the standardized features. The first two principal components (PC1 and PC2) capture ~67.6% of the total variance (PC1: 42.4%, PC2: 25.2%). The component loadings interpret as follows:

- **PC1 (Price Axis):** Strong positive loadings on log-offer price (0.63) and log-real price (0.62), with minor contributions from ratings/reviews. This axis primarily represents an overall price **level**. Products with high PC1 scores are higher-priced; those with low PC1 scores are low-priced.
- **PC2 (Engagement/Quality Axis):** Strong positive loadings on log-review count (0.72) and ratings (0.52), with small negative contributions from prices. This axis captures consumer engagement and satisfaction. High PC2 means many reviews and high ratings; low PC2 means low popularity and/or ratings.

In the PC1–PC2 scatterplot, clusters partially separate:

- **Cluster 0** points are mostly on the lower side of PC1 (lower price) and moderate on PC2, reflecting affordable products with average engagement.
- **Cluster 1** points shift to the right (higher PC1) and moderate-to-high on PC2, reflecting expensive products with strong engagement.
- **Cluster 2** points are around mid-range PC1 but lower on PC2, showing mid-priced products with low engagement.
- **Cluster 3** points are widely scattered, often to the high end of PC1 and low end of PC2, indicating expensive products with poor engagement.

The clusters overlap, confirming the "soft" segmentation observed earlier. PCA also quantifies feature importance: people_review_log has the highest influence (as seen in PC2), followed by price variables; ratings contribute less than these.

The PCA analysis aligns with our earlier findings: price and popularity emerge as the main differentiators, whereas ratings play a secondary role. Products compete mainly on pricing and visibility, not ratings.

Results

The clustering and analyses consistently identify two product groups with **relatively weaker observable market traction**:

1. Cluster 2 (Higher-Rated, Low-Engagement Products)

- **Characteristics:**
  Above-average prices with the highest average user ratings (≈ 4.24) and an extremely low average cumulative review count (≈ 312). Price and rating variances are moderate, while review counts are uniformly low across the cluster.
- **Interpretation:**
  This cluster represents products that receive relatively favorable ratings from a small number of reviewers but exhibit consistently low observable engagement. The signal configuration indicates weak observable market traction driven by low cumulative review volume rather than low ratings.
- **Implication:**
  Products in this cluster warrant closer monitoring to assess whether engagement levels change over time. Further investigation may help determine whether low observed engagement persists or evolves as additional market information becomes available.

2. Cluster 3 (Higher-Priced, Low-Engagement Products with Greater Dispersion)

- **Characteristics:**
  The highest average prices (offer ≈ ₹1,248; real ≈ ₹5,113), the lowest average ratings (≈ 3.55), and low average cumulative review counts (≈ 1,107). This cluster exhibits high variance in both prices and ratings.
- **Interpretation:**
  This cluster combines higher price levels with lower ratings and limited engagement, along with

substantial internal heterogeneity. The observable signal pattern reflects another distinct form of weak market traction within the dataset.

- **Implication:**
  Products in this cluster may be prioritized for further examination to understand whether these observable signal patterns remain stable or change over time.

For completeness, the remaining clusters exhibit comparatively stronger observable market traction:

- **Cluster 0 (Lower-Priced, Moderately Engaged Products):**
  Lower prices (offer ≈ ₹561), moderate average ratings (≈ 3.79), and relatively higher cumulative review counts. These products display comparatively stronger observable engagement and are not primary candidates for closer monitoring.
- **Cluster 1 (Higher-Priced, Highly Engaged Products):**
  Higher prices (offer ≈ ₹1,198), stable average ratings (≈ 3.98), and the highest cumulative review counts (≈ 152k). This cluster reflects the strongest observable market traction within the dataset and is of lower priority for monitoring in the present analysis.

In summary, products in **Clusters 2 and 3** exhibit relatively weaker observable market traction based on cumulative engagement patterns, while **Clusters 0 and 1** display comparatively stronger observable traction and are not the primary focus of this study.

Strategic Recommendations

Based on these insights:

- **Cluster 2 (Hypotheses for Growth-Oriented Experiments): Products** in this cluster combine relatively higher ratings with low cumulative engagement, suggesting they may be candidates for testing visibility-oriented interventions. Possible experiments include temporarily featuring a subset more prominently on the platform, running targeted promotions, or bundling them with already popular products to observe whether review accumulation and other observable signals change over time. Any uplift would need to be evaluated empirically rather than assumed.
- **Cluster 3 (Hypotheses for Diagnostic Review):** Products in this cluster combine higher prices with lower ratings and limited engagement, together with relatively high internal dispersion. These characteristics indicate candidates for more detailed diagnostic review, such as examining customer feedback, comparing pricing and feature sets with competing products, or monitoring a sample of items for changes in observable signals following adjustments. Decisions about continuation, repricing, or repositioning should be made at the individual product level and supported by additional evidence.
- **General Monitoring:** Across all clusters, ongoing tracking of review accumulation and rating dynamics would help distinguish persistent low-traction patterns from early lifecycle effects, especially given the absence of product age and exposure data. Incorporating time-series information in future work would allow more precise differentiation between newly launched products and structurally underperforming offerings.

Limitations

This analysis relies exclusively on publicly observable market signals, namely price, user ratings, and cumulative review volume, which serve as proxies rather than direct measures of market performance. Product age, exposure, and sales data are unavailable; therefore, low observed traction may reflect early lifecycle stages or limited visibility rather than true underperformance. Review counts represent cumulative engagement and do not capture recent growth trends or momentum. Additionally, user ratings may be affected by self-selection bias, as feedback is provided by a non-random subset of consumers. Finally, the use of distance-based clustering methods imposes geometric assumptions that may oversimplify complex market behavior. Consequently, the resulting clusters should be interpreted as relative intervention priority groups rather than definitive product lifecycle or performance classifications.

Future Scope

Future work may incorporate product age, exposure, and time-series review trends to distinguish early-stage products from structural underperformers. Integrating sales and conversion metrics, expanding feature sets, and developing predictive models could enable proactive intervention forecasting.

Conclusion

This study applies K-Means and agglomerative hierarchical clustering to segment true wireless earphones on Flipkart using offer price, real price, cumulative review volume, and user ratings, with the goal of highlighting products that exhibit relatively weaker observable traction signals under public-data constraints. The analysis shows that traction is multidimensional and continuous: pairwise feature-space views, cluster-wise summaries, and PCA all reveal substantial overlap between clusters, so no single variable can summarize market behavior on its own.

Within this structure, two clusters (2 and 3) stand out for weaker observable traction, both marked by low cumulative engagement but differing in their combinations of ratings and price levels—one with relatively higher ratings, the other with higher prices and lower ratings. This indicates that weak observable traction is not a single pattern and cannot be inferred from price or ratings in isolation. These clusters are best interpreted as relative monitoring and investigation priority groups rather than as definitive evidence about product quality, exposure, or commercial outcomes, and any causal or outcome-based claims remain beyond the scope of the available signals. Overall, the work demonstrates how unsupervised learning can structure limited observable market data into interpretable patterns to flag products for closer review, while emphasizing the need for cautious, context-aware interpretation in real e-commerce settings.