UNCOVERING THEMES IN AMAZON REVIEWS USING TEXT ANALYSIS

BATSHEVA LEVIN

Agenda

- Business Problem & Objective
- Business Value
- Data Overview & EDA
- Methodology & Model Selection
- Clustering Results & Insights
- Business Recommendations
- Conclusion

Business Problem

Amazon receives millions of product reviews, making it difficult to manually analyze customer feedback.

There is a need to understand what customers like and dislike, and how products and services can be improved.

Objective

Apply Natural Language Processing (NLP) and unsupervised clustering to group similar reviews.

Identify key themes, sentiment patterns, and customer concerns to develop practical business recommendations.

Business Value

Helps Amazon and third-party sellers identify issues related to products, shipping, or customer service

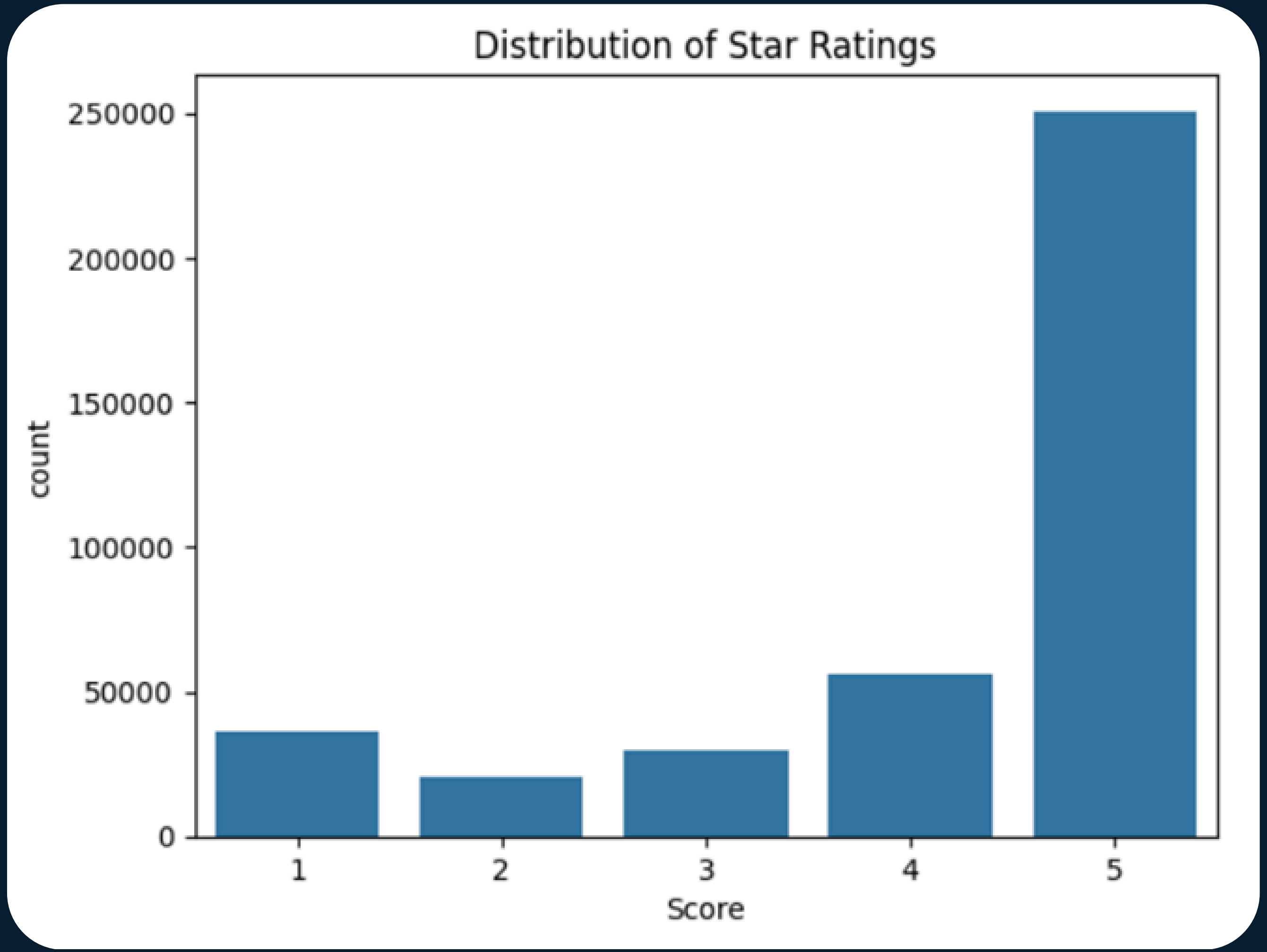
Provides actionable insights to improve products, marketing, and service

Enhances overall customer satisfaction

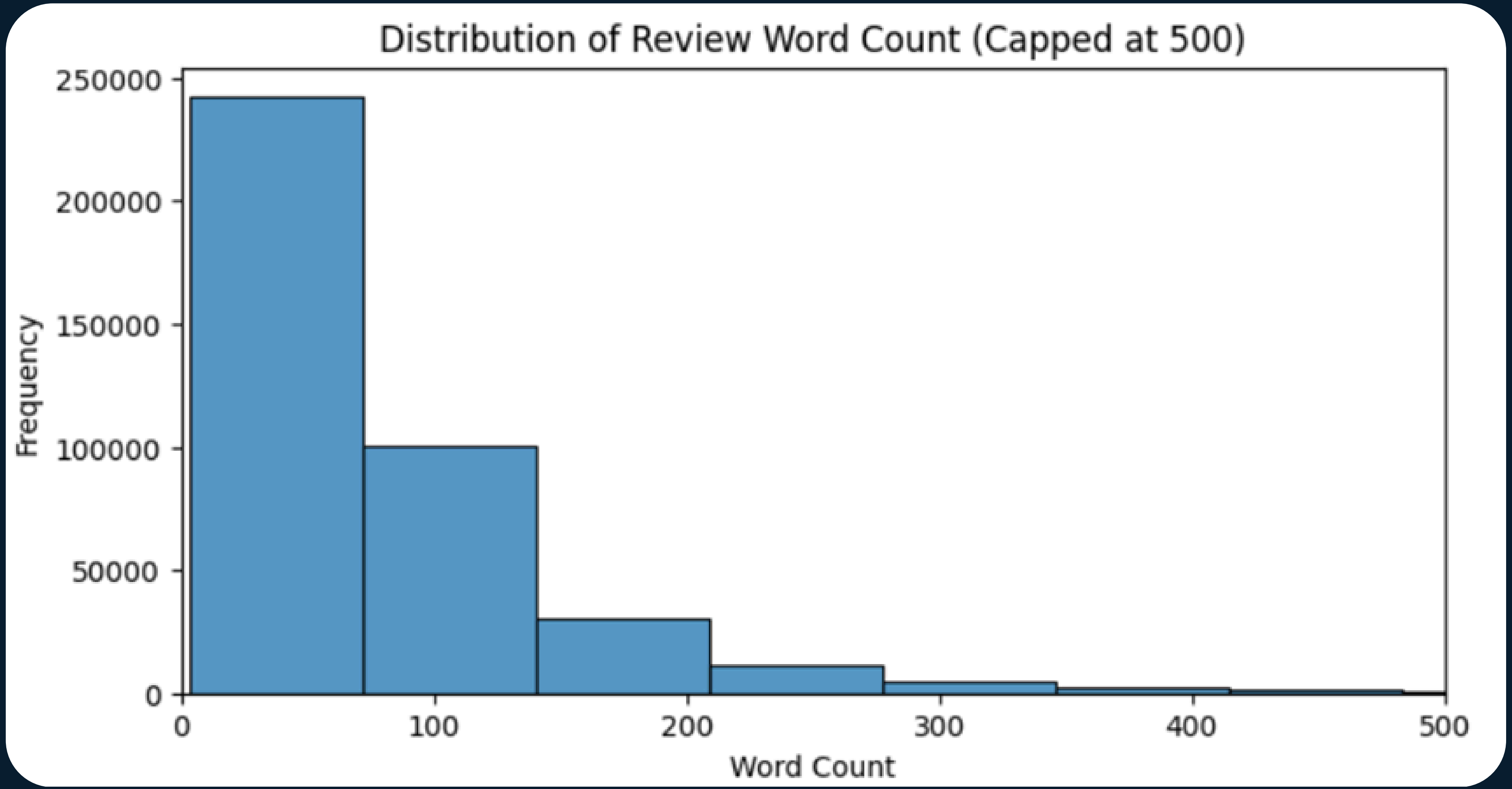
Data Overview

- Approximately 390,000 Amazon reviews
- Time range: October 8, 1999 – October 26, 2012
- Reviews focused on food products: drinks, snacks, cooking ingredients, pet treats, etc.
- Mainly used the ‘Text’ field for analysis
- Star ratings were used later for evaluation
- Unsupervised learning approach - no traditional target or predictors

Exploratory Data Analysis



Star Rating Distribution: Skewed toward 5-star reviews, indicating class imbalance



Review Lengths: Most reviews are less than 100 reviews, with some outliers exceeding 500

Methodology/Approach

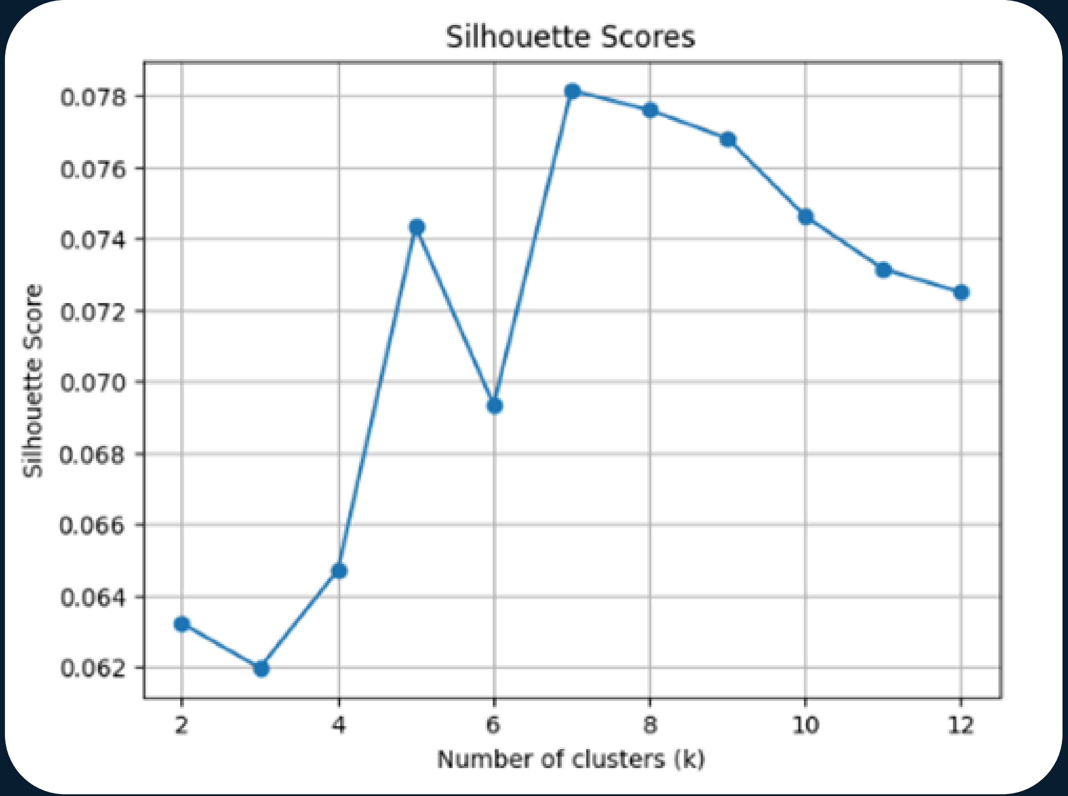
- **Preprocessing**: Cleaned review text by getting rid of duplicates, converting text to lowercase, and removing non-alphanumeric characters
- **Sentence-BERT Embeddings:** Generated embedding vectors to represent the semantic meaning of each review
- **Dimensionality reduction:** Applied StandardScaler to standardize the embeddings and used PCA to reduce dimensionality from 384 to 50 for better clustering efficiency
- **Clustering**: Applied K-means to group reviews based on similarity

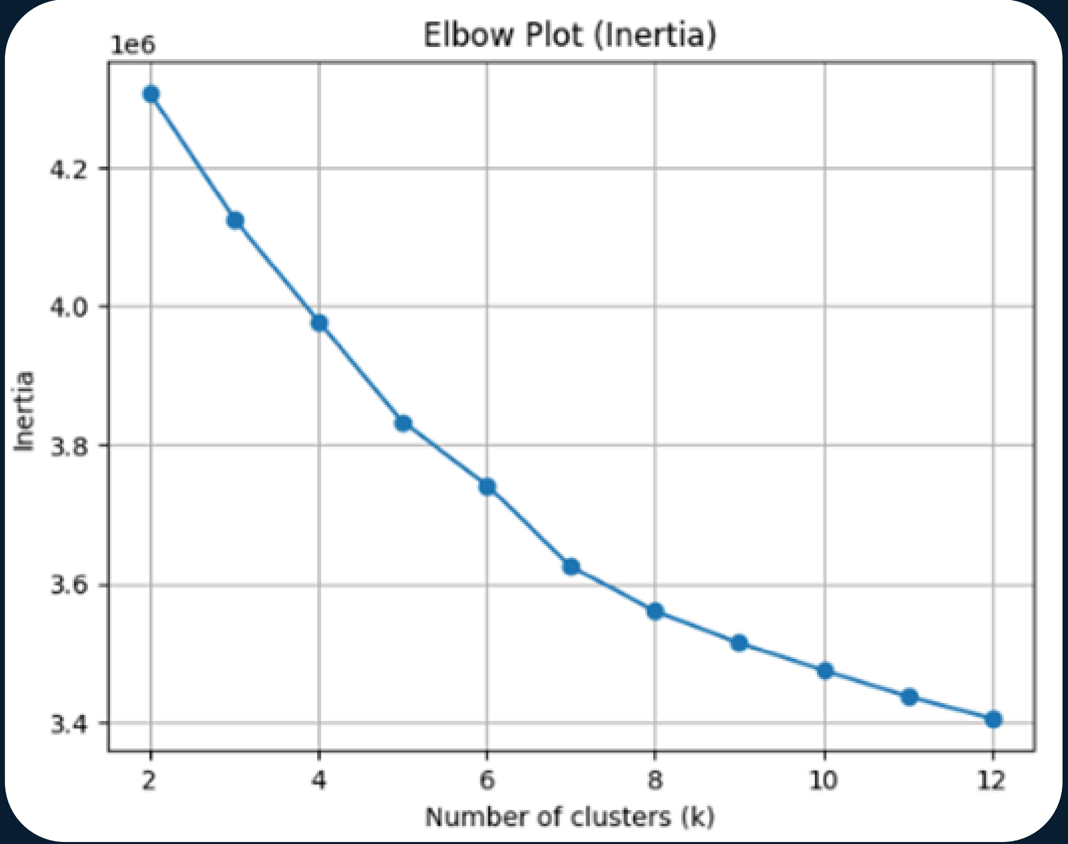
Model Selection

Tested a range of *k* values (number of clusters) on a sample of the data

Evaluated using both the silhouette score and the elbow method (inertia)

Final *k* = 7 selected based on these two metrics



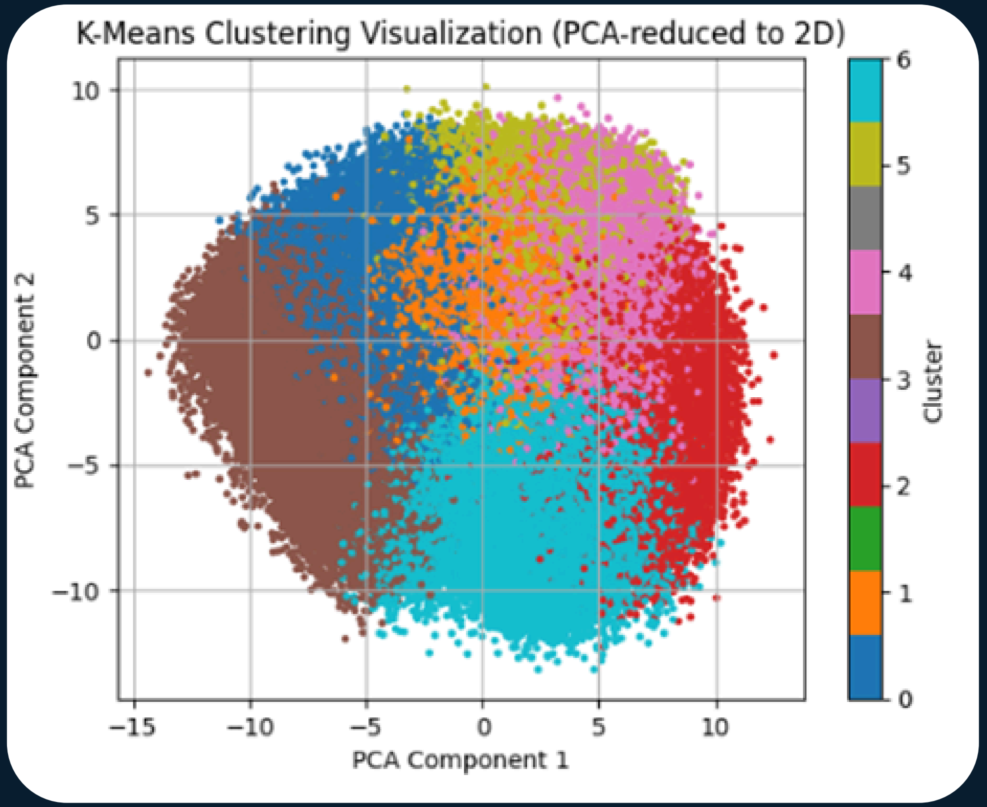


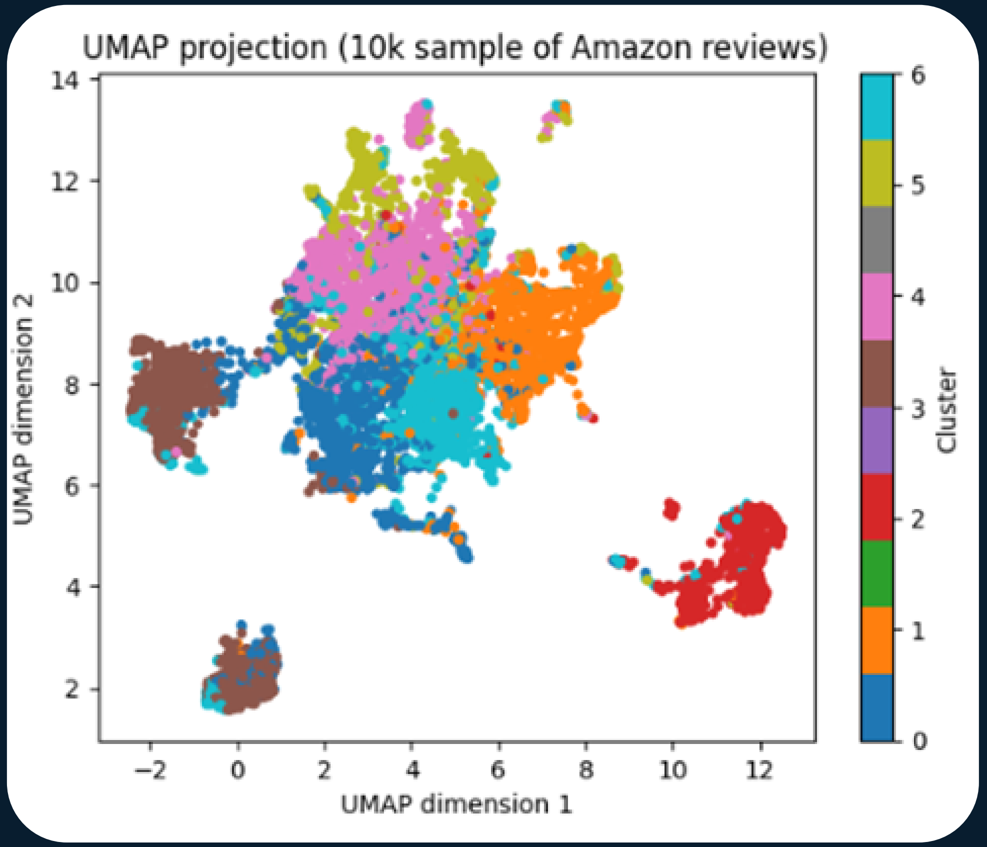
Cluster Visualization

Visualized clusters using both UMAP and PCA (reduced to 2D)

Plots show some overlap between clusters

Despite the visual overlap, the reviews within each cluster were thematically similar





Sample Reviews & Themes by Cluster

For each cluster, the reviews closest to the cluster centroid were examined to understand the major themes. Top keywords were extracted using TF-IDF, a method for identifying important words in text.

Cluster	Theme	Top TF-IDF Keywords	Sample Review Excerpts
0	Natural Sweeteners & Drink Mixes	drink, flavor, taste, sweet	“Best tasting sugar-free punch.” “Did not taste good… small portion.”
1	Flavorful Cooking Ingredients	flavor, food, sauce, taste	“Great flavor, filling, healthy.” “Super convenient… but way too salty.”
2	Specialty Pet Food & Treats	cat, dog, food, treat	“This is the only food my picky dog eats.” “Our dog loves it… and our vet approves”

Sample Reviews & Themes by Cluster (cont.)

Cluster	Theme	Top TF-IDF Keywords	Sample Review Excerpts
3	Premium Coffee Products	coffee, cup, drink, strong	“Highly recommend Britt Coffee.” “….weak and almost tasteless.”
4	Sweet Snacks and Candy	candy, chips, snack, sweet	“Amazing candies… must try!” “My go-to snack… just wish it was healthier.”
5	Healthy Breakfast and Snacks	bread, cereal, gluten, taste	“…Healthy and tastes good.” “Delicious breakfast… But too much added sugar.”
6	General Food Purchases and Experience	arrived, box, order, price, shipping	“Delicious… but a bit pricey.” “…Old, dry and overpriced."

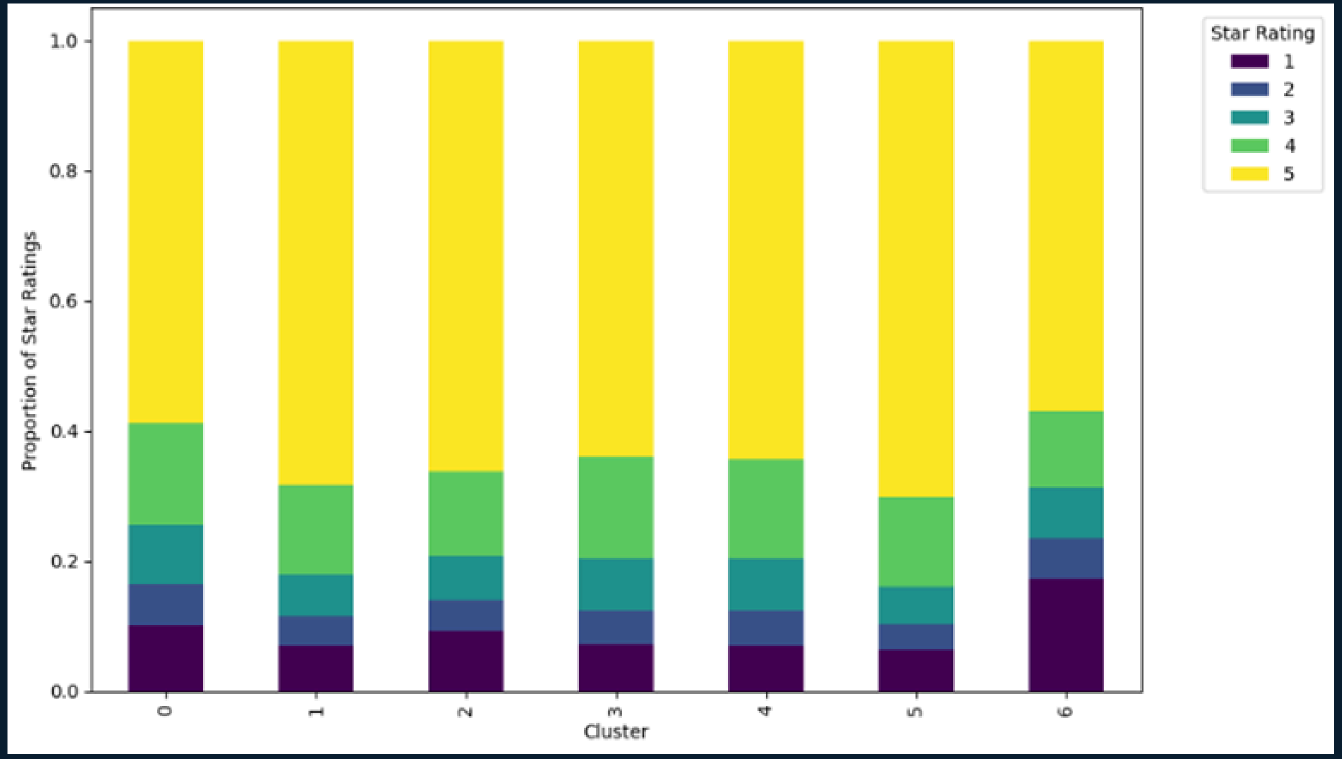
Star Rating Analysis

Number of Reviews per Cluster

Cluster	Number of Reviews
0	63,530
1	60,515
2	39,310
3	64,886
4	66,293
5	40,549
6	58,496

Average Star Rating by Cluster

Cluster	Average Star Rating
0	4.07
1	4.32
2	4.22
3	4.24
4	4.25
5	4.37
6	3.85



Stacked bar chart showing how star ratings are distributed within each cluster.

Cluster Insights and Business Recommendations

Cluster	Topic	Business Recommendation
0	Natural Sweeteners & Drink Mixes	Offer value packs or larger sizes to address complaints about small portions.
1	Flavorful Cooking Ingredients	Share recipes or cooking tips to inspire new product uses.
2	Specialty Pet Food & Treats	Emphasize vet approval and high-quality ingredients to build customer trust.
3	Premium Coffee Products	Clarify flavor profiles on packaging (strong, mild, sweet, etc.) to set expectations.
4	Sweet Snacks and Candy	Introduce low-sugar options to appeal to a wider range of preferences.
5	Healthy Breakfast and Snacks	Highlight nutritional benefits (fiber, protein, whole grains, etc.) on packaging.
6	General Food Purchases and Experience	Improve packaging and freshness to reduce complaints about quality upon arrival.

Conclusion

In this project, hundreds of thousands of Amazon food product reviews were analyzed to uncover what truly matters to customers. By examining the review text, we gained deeper insights that star ratings alone could not provide. Clustering similar reviews revealed seven distinct themes, each highlighting specific customer preferences and concerns. These findings can help Amazon and its sellers improve products, refine marketing strategies, and ultimately drive more sales while creating a better customer experience.