Market Segmentation of Agricultural Markets in India Using PCA and K-Means Clustering

Data Science Team

Ritesh Singh

June 23,2025

Table of Contents

1. Introduction

1.1 Objective:

The goal of this analysis is to identify distinct market segments across Indian markets using clustering techniques. This enables a better understanding of price behaviour, commodity diversity, and regional patterns, which are crucial for policy formulation, procurement strategy, and infrastructure planning.

1.2 **Motivation:**
Market dynamics in agriculture vary widely, from high-value niche markets to price-sensitive, broad-distribution zones. Segmentation helps decode this complexity

2. Data Preprocessing and Outlier Removal

2.1 Dataset Summary:

Aggregated mandi-level data across Indian districts, including features such as modal price, price volatility, and commodity/variety diversity.

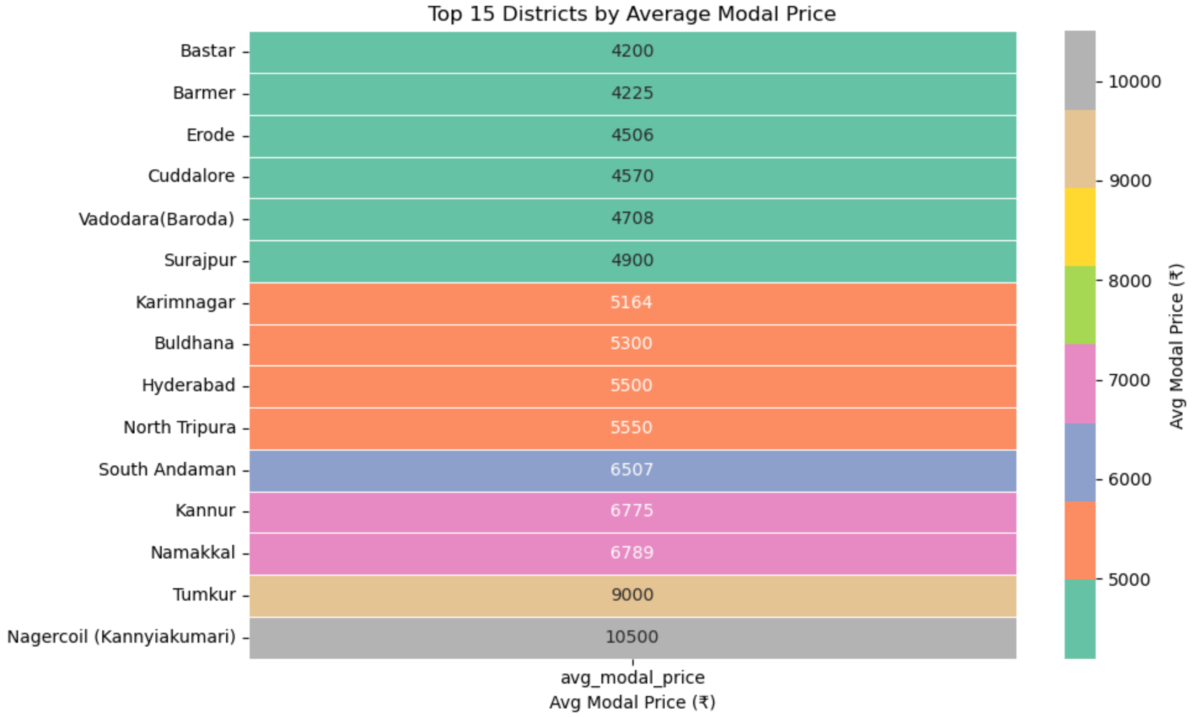
2.2 Data Cleaning:
Removed two duplicates, removed rows which contained zero in min_price and max_price.

2.2 Grouping Data:
Grouped the cleaned data by state, district, market to get ['state', 'district', 'market', 'unique_commodities', 'unique_varieties', 'avg_modal_price', 'std_modal_price', 'avg_max_price', 'avg_min_price']

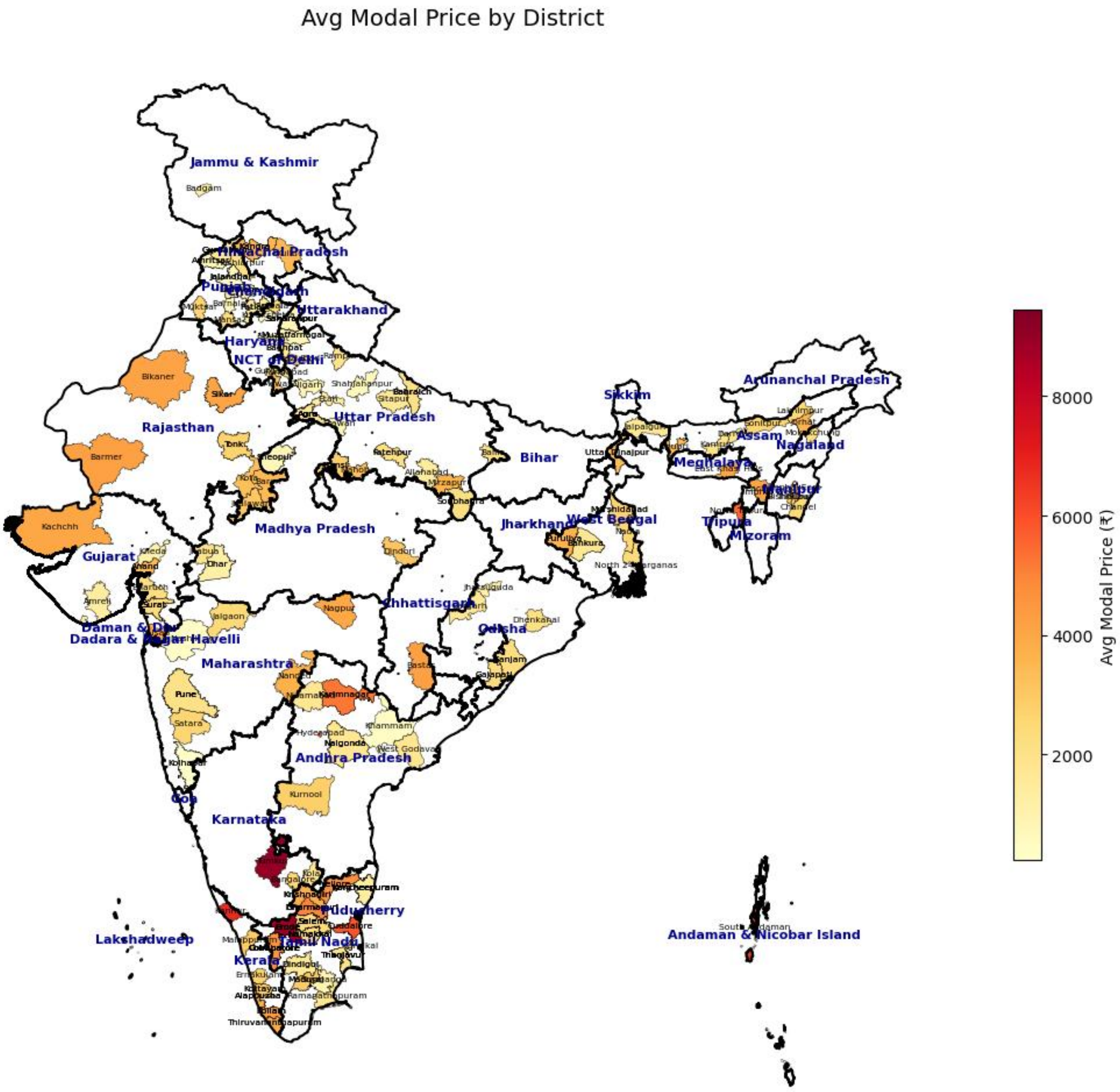
2.3 **Outlier Filtering:**
Prices beyond the 1st and 99th percentiles were removed to minimize distortion caused by extreme values. Two datasets were maintained:

- market_summary: Cleaned, outlier-removed data used for clustering
- market_summary_with_outliers: Full dataset for comparison and validation

3. Visualization Before Segmentation



- Nagercoil (Kannyiakumari) are Tumkur are the highest-priced markets in the snapshot, with average modal prices above ₹10500, ₹9000. These are potentially high-value markets
- Bater, Barmer and erode are at the lower end, averaging around ₹4500. These might be smaller markets.
- In our context—evaluating markets for a smart crop disease detection app—districts with higher modal prices might be more open to adopting technology, since their crops could carry more financial risk if disease strikes.



- Tamil Nadu and Panjab have lots of districts with good avg modal price
- Goa and South Andaman have top avg modal price

4. Dimensionality Reduction with PCA

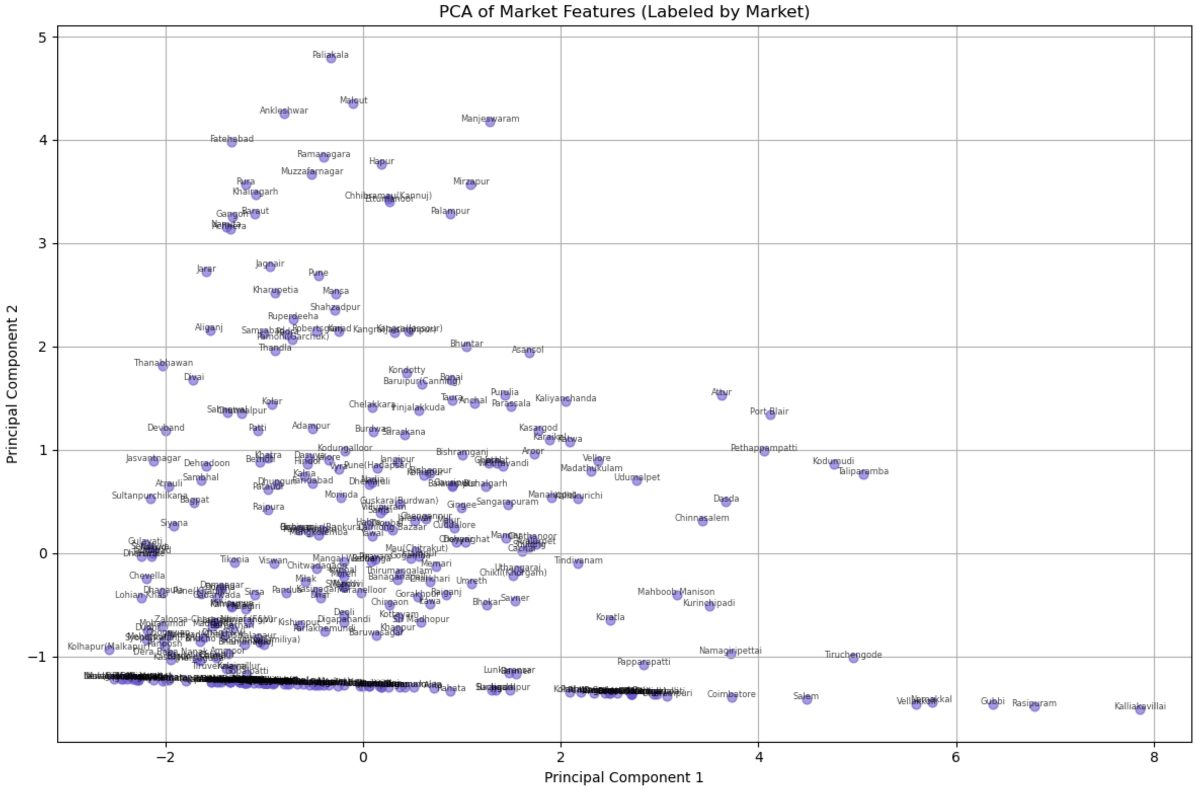
4.1 Goal:

Reduce feature space to 2 principal components while preserving key variation.

4.2 Outcome:

PCA revealed clear separation among markets based on:

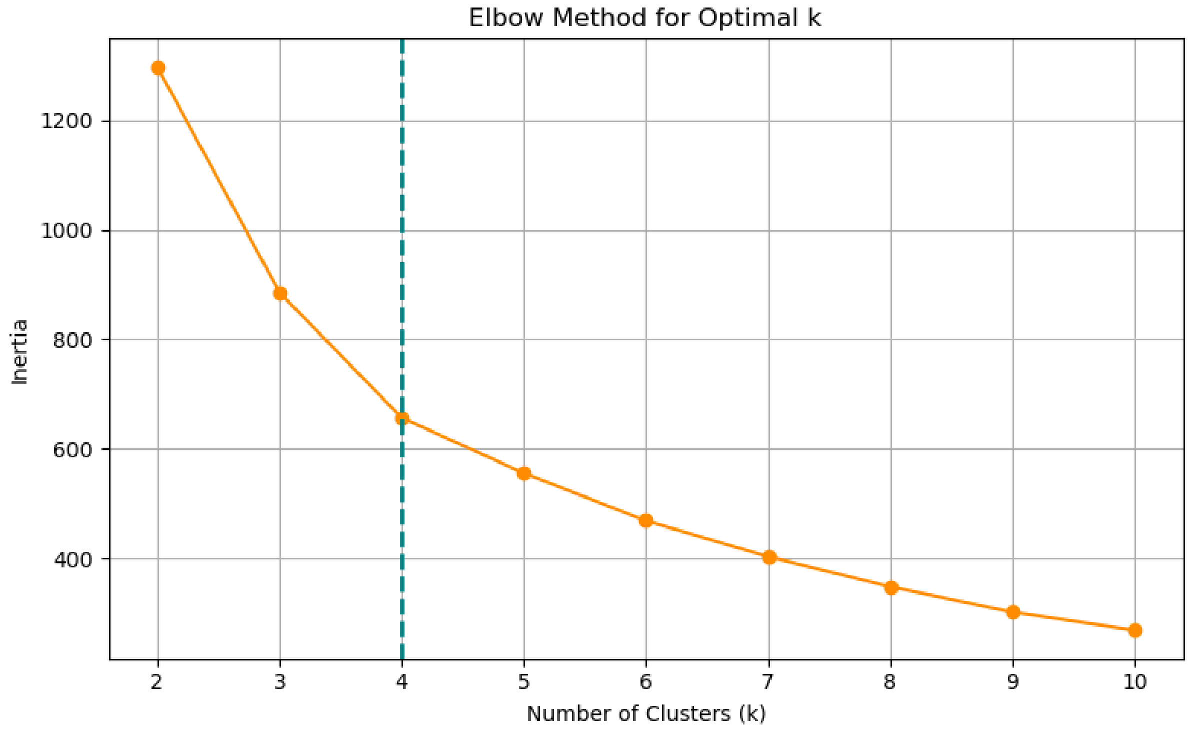
- Pricing behaviour (avg and std of modal price)
- Diversity in crops and varieties.



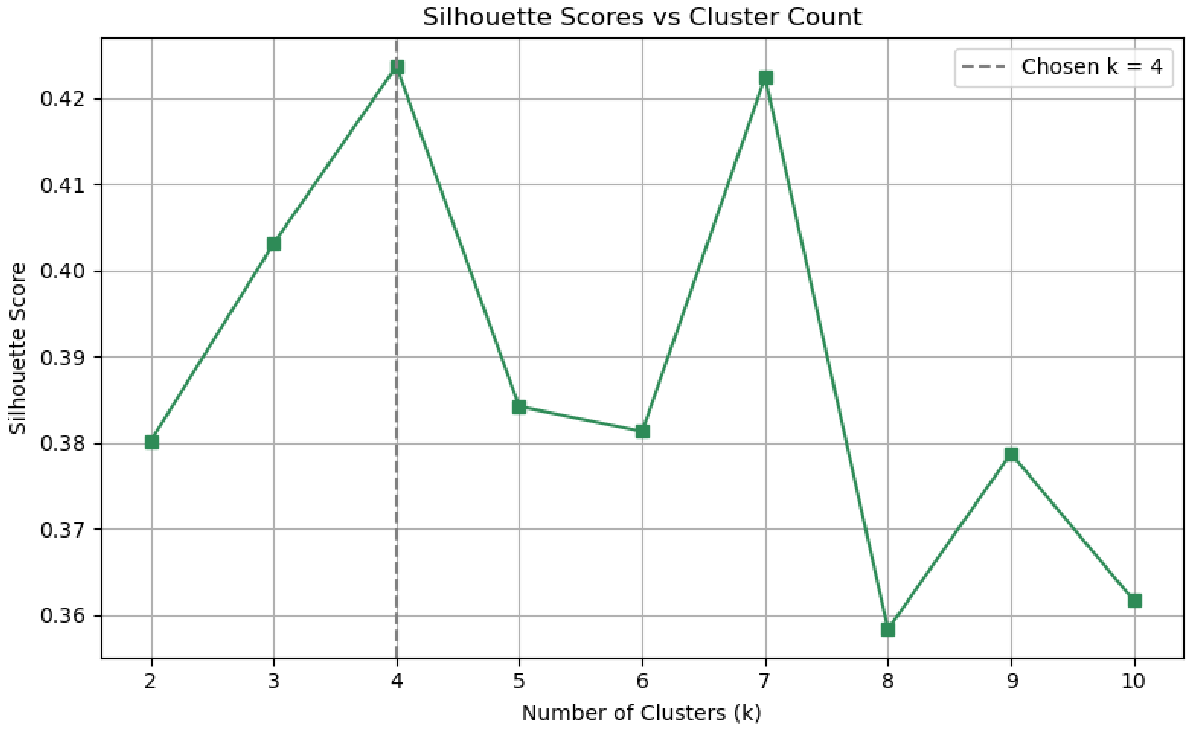
5. Cluster Selection and Segmentation

Methods Used:

5.1 **Elbow Method:** Inertia curve flattened after 4 clusters

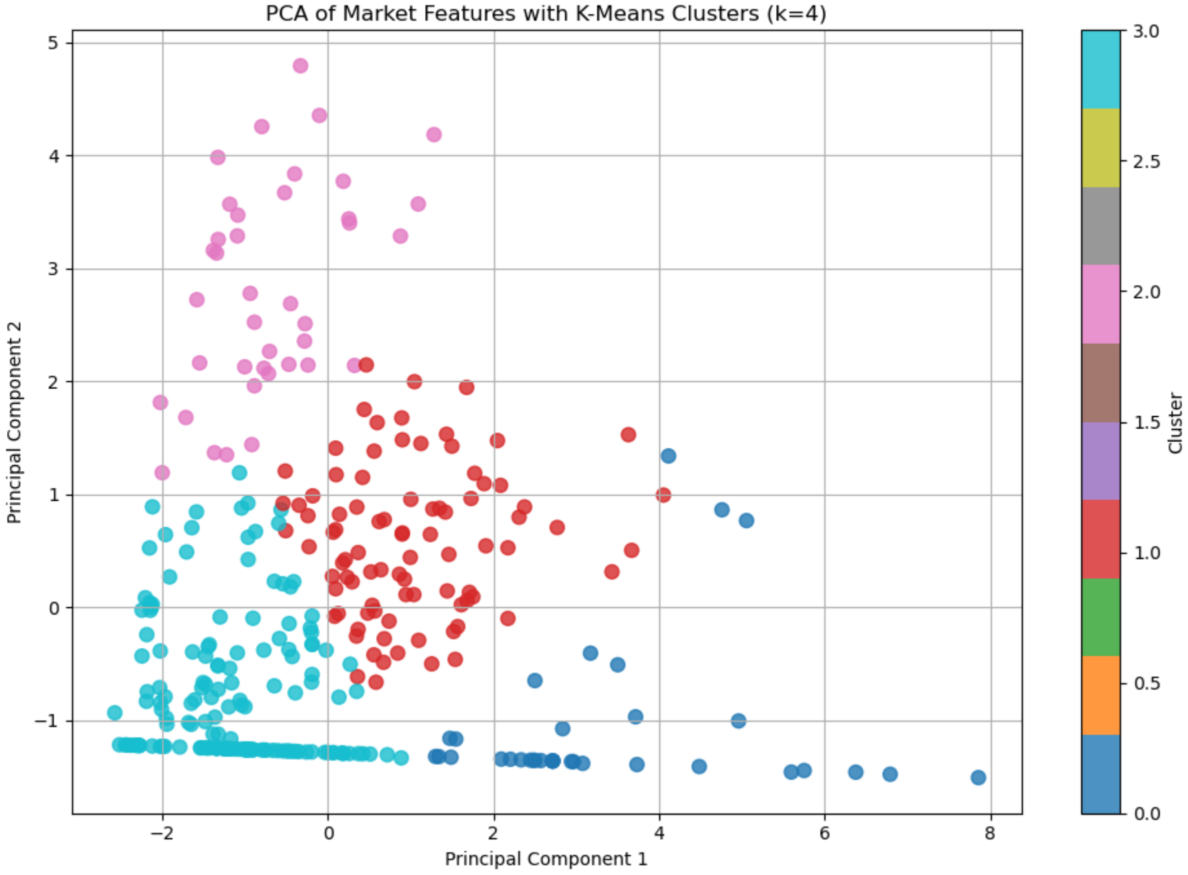


5.2 **Silhouette Analysis:** Peak silhouette score at k = 4



6. K-Means Clustering and Interpretation

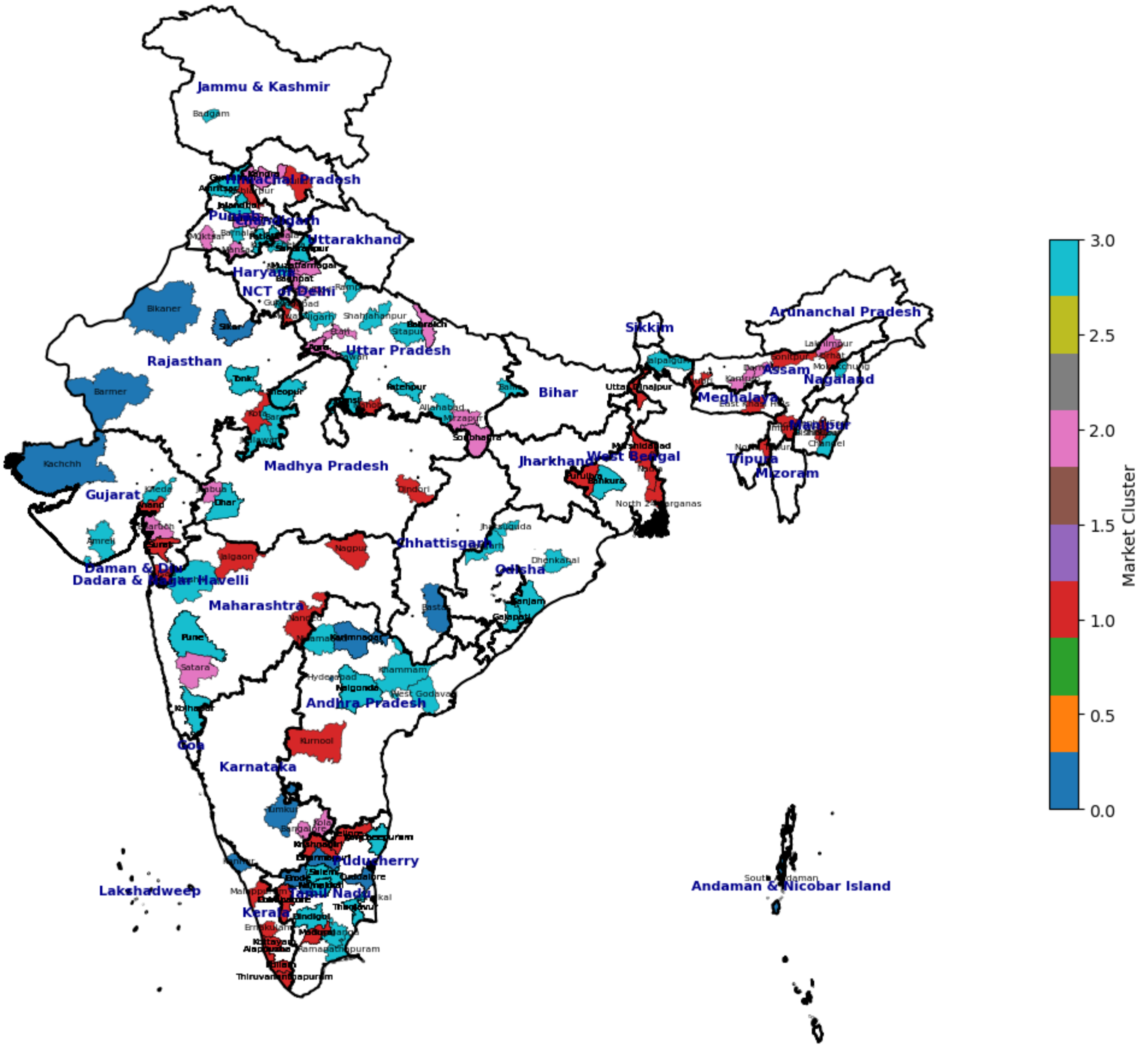
6.1 PCA of Market Features with K-Means Clusters (k=4):



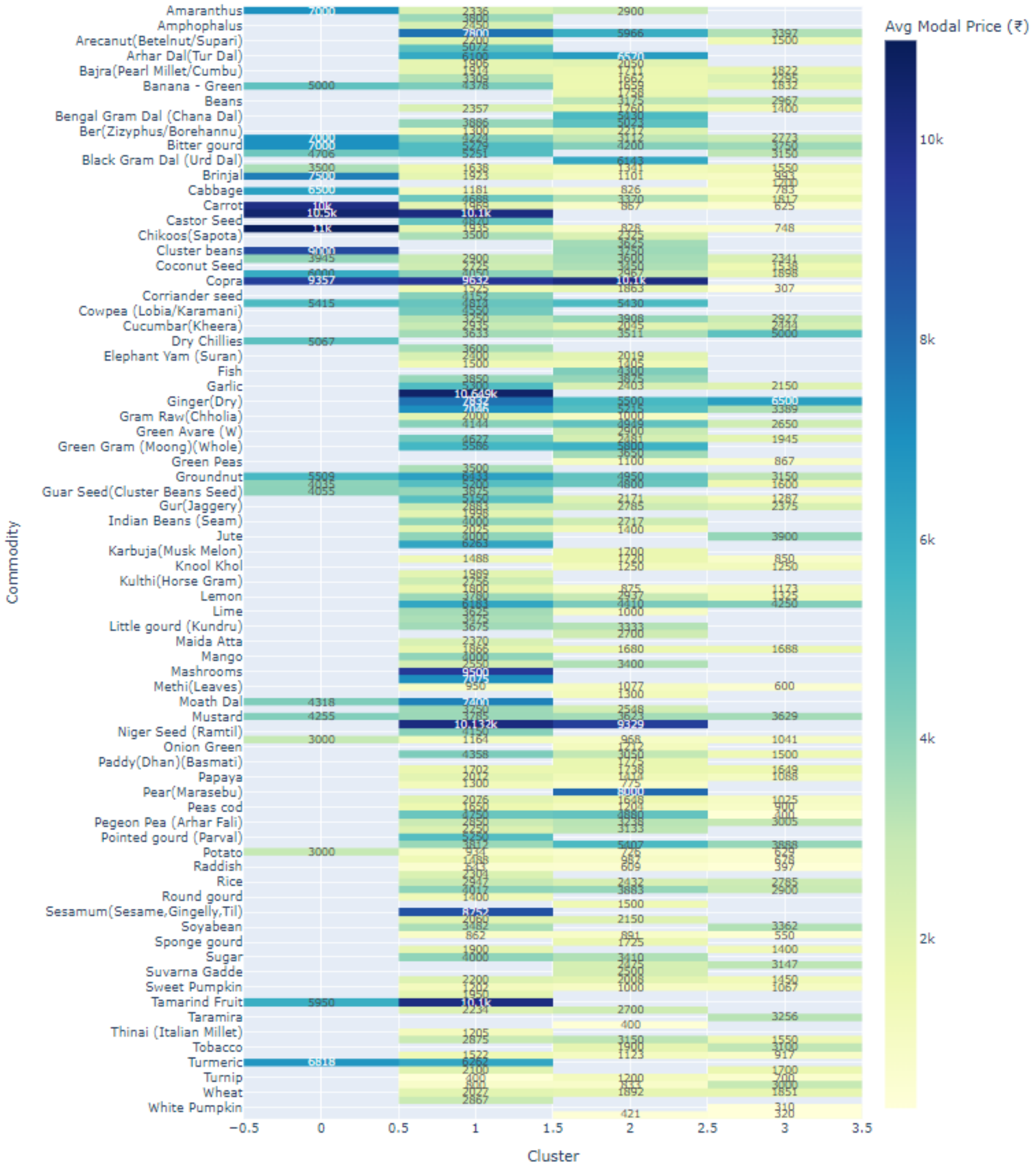
6.2 Geospatial Distribution of Clusters

- **Approach:**
  Merged cluster labels into district-level shapefiles for mapping
- **Insights:**
  - Certain high-price clusters concentrate in coastal or island regions
  - Diverse markets are spread across agriculturally active states
  - Volatile clusters show wide regional dispersion

Market Clusters by District (K-Means = 4)



6.3 Cluster vs Commodity Heatmap:



6.4 Profiling Clusters

cluster	**unique_commodities**	**unique_varieties**	**avg_modal_price**	**std_modal_price**	**avg_max_price**	**avg_min_price**
0	1.6	1.4	6013.4	558.8	6375.5	5570.5
1	8.2	3.7	3367.1	2282.6	3538.1	3163.2
2	19.8	15.9	2093.9	1583.5	2225.6	1933.6
3	3.3	2.2	1602.2	375.6	1735.2	1450.7

Clusters	Profile Summary	Suggested Name
0	Very low diversity, high prices, stable	Premium, Focused Markets
1	Moderate diversity, high volatility	Volatile Multi-Crop Hubs
2	High diversity, moderate pricing, moderate volatility	Diverse & Dynamic Markets
3	Low diversity, lowest prices, stable	Affordable, Low-Variety Zones

7. Conclusion and Future Scope

Conclusion: The clustering successfully revealed distinct market segments with actionable characteristics. These segments can guide policy, procurement, and supply chain optimization.