Crop Yield Analysis in India

Executive Summary

This report presents a comprehensive analysis of crop yields in India based on a dataset containing 19,689 records spanning from 1997 to 2020. The analysis explores patterns and relationships to identify factors that influence crop yields and provides data-driven recommendations for farmers.

Key Findings:

- Production shows the strongest correlation with yield (correlation coefficient: 0.57)
- Coconut, Sugarcane, and Banana are the highest-yielding crops in India
- Five distinct farm clusters were identified based on area, rainfall, fertilizer, and pesticide usage
- Different crops perform optimally under specific conditions of rainfall, season, and agricultural inputs

1. Data Exploration

The analysis was conducted on a dataset containing 19689 records of crop production in India. The dataset includes 55 unique crops grown across 30 states during 6 different seasons, spanning from 1997 to 2020.

1.1 Yield Distribution

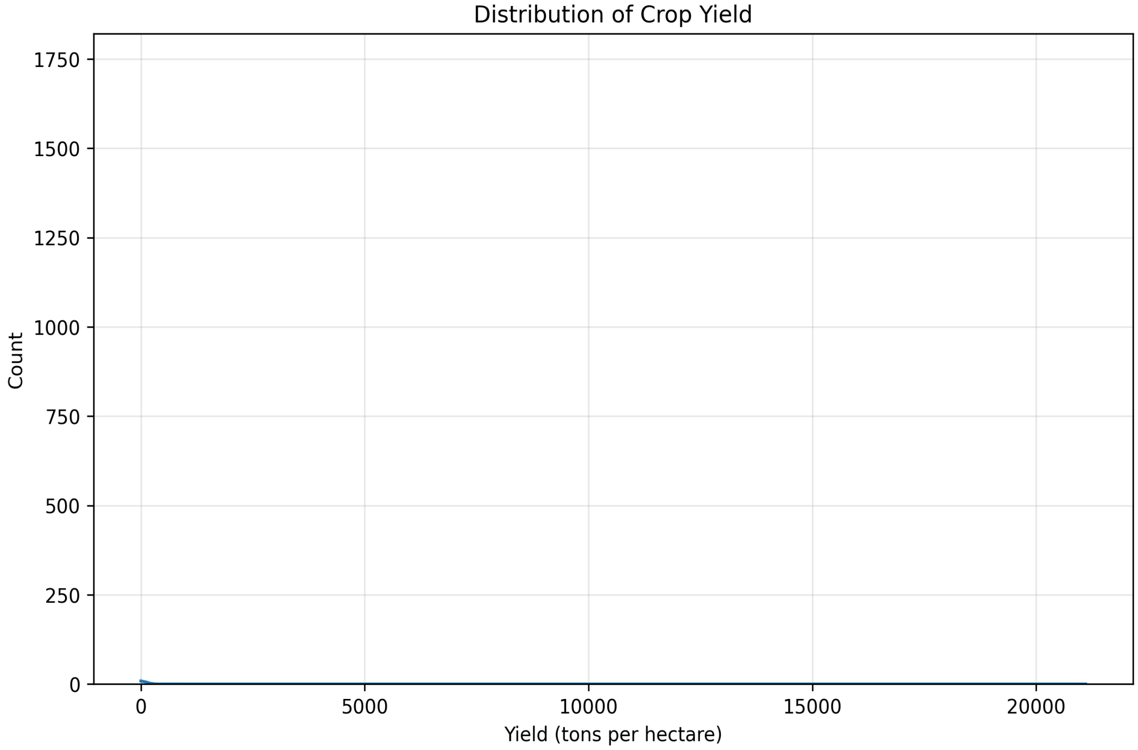

Figure 1: Distribution of crop yields across all data points

Observations: The yield distribution shows a significant right skew, indicating that most crops have relatively low yields, while a small number of crops have exceptionally high yields. Coconut stands out with yields orders of magnitude higher than other crops, which reflects its production measurement in nuts per hectare rather than weight. The long tail of the distribution suggests that specific high-yielding crops like Coconut, Sugarcane, and Banana significantly outperform other crops in terms of yield.

1.2 Yield by Crop and Season

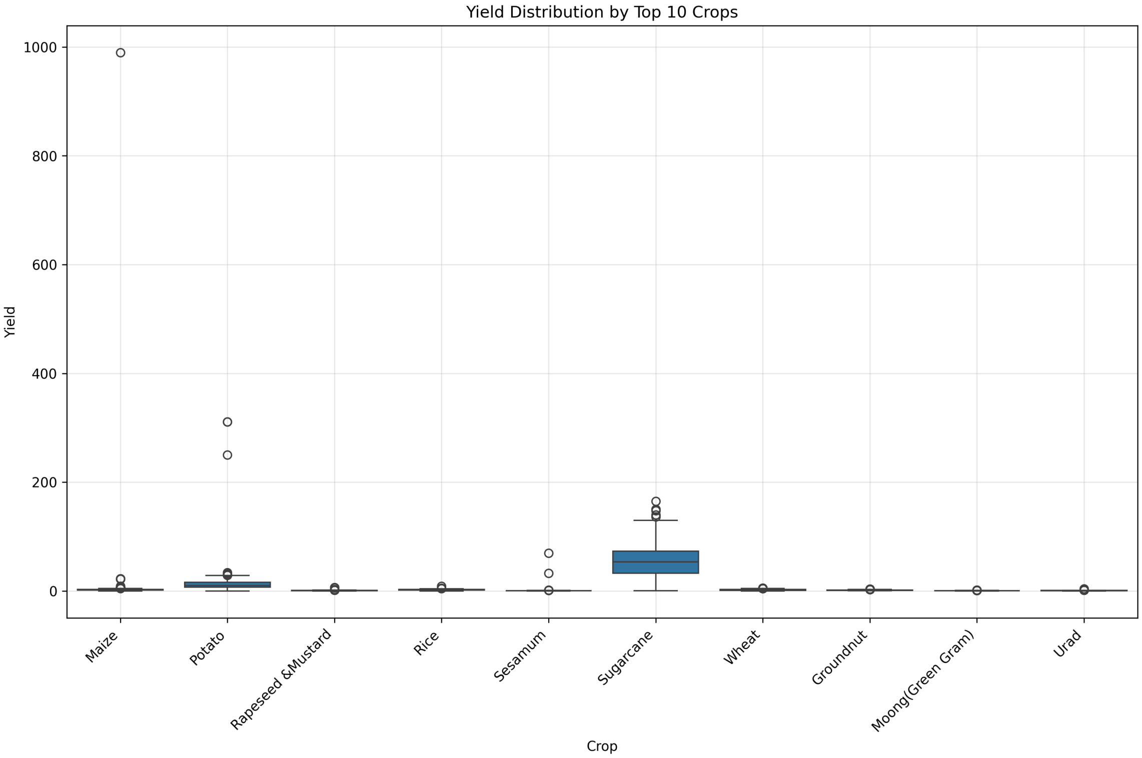

Figure 2: Yield distribution across top 10 crops

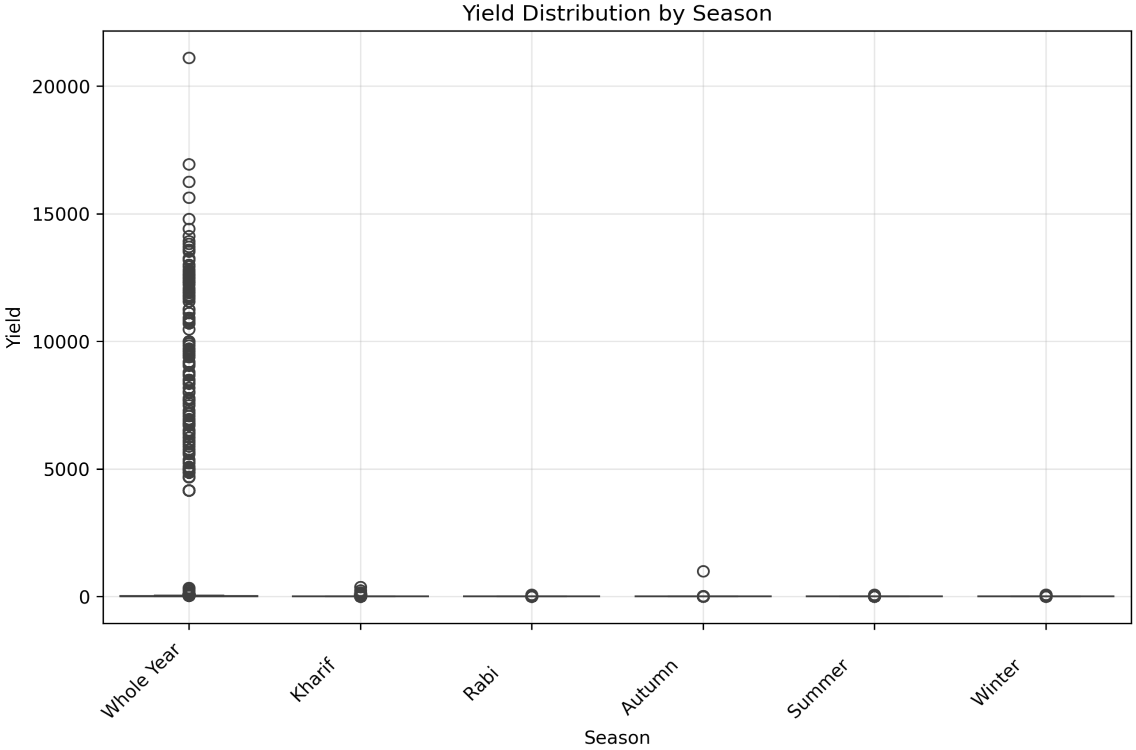

Figure 3: Yield distribution across different seasons

Observations on Crop Yields: The boxplot of top 10 crops reveals extreme variation in yields across different crop types. Coconut shows the highest median yield by a significant margin, followed by Sugarcane and Banana. These three crops consistently outperform others, making them potentially lucrative choices for farmers. The wide range and presence of outliers in some crops like Sugarcane indicate that yield can vary considerably depending on growing conditions and farming practices. **Observations on Seasonal Yields:** The seasonal analysis shows that crops grown during the "Whole Year" tend to have higher median yields compared to crops grown in specific seasons. "Summer" crops generally show higher yields than "Winter" or "Kharif" (monsoon) season crops, which could be attributed to better growing conditions or the types of crops typically grown in each season. This suggests that farmers might benefit from selecting crops that can be cultivated throughout the year or during summer when possible.

1.3 Top Performing Crops

Crop	Average Yield
Coconut	8652.00
Sugarcane	51.73
Banana	26.85
Tapioca	16.67
Potato	13.33

Observations on Top Crops: The top five crops by average yield show a stark contrast in productivity. Coconut leads with an extraordinary average yield of over 8,600 units per hectare, primarily because it's measured in nuts rather than weight. Sugarcane follows with approximately 52 tons per hectare, making it one of the most productive crops by weight. Banana, Tapioca, and Potato round out the top five with yields of 27, 17, and 13 tons per hectare respectively. These high-yielding crops represent potential opportunities for farmers looking to maximize productivity per unit of land.

2. Correlation Analysis

A correlation analysis was performed to identify relationships between different variables and crop yield.

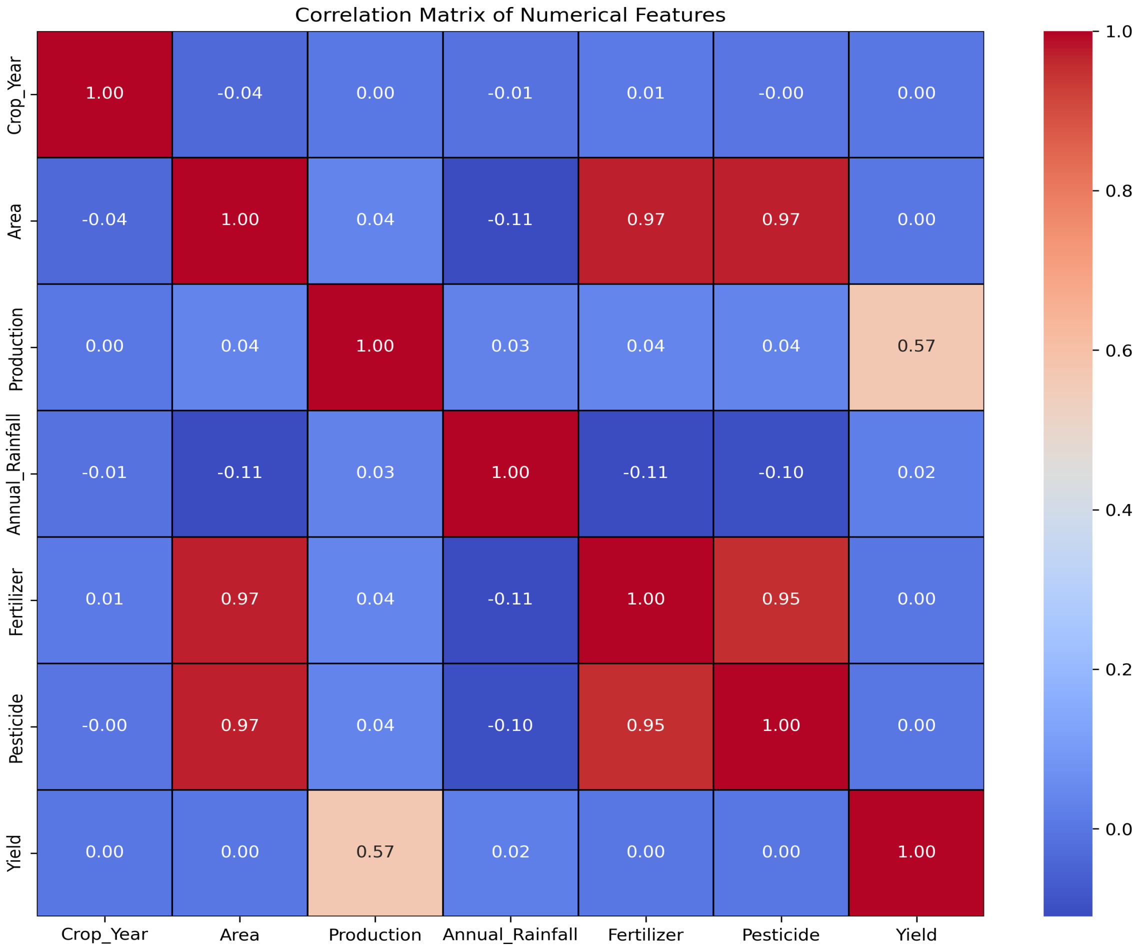

Figure 4: Correlation matrix showing relationships between different variables

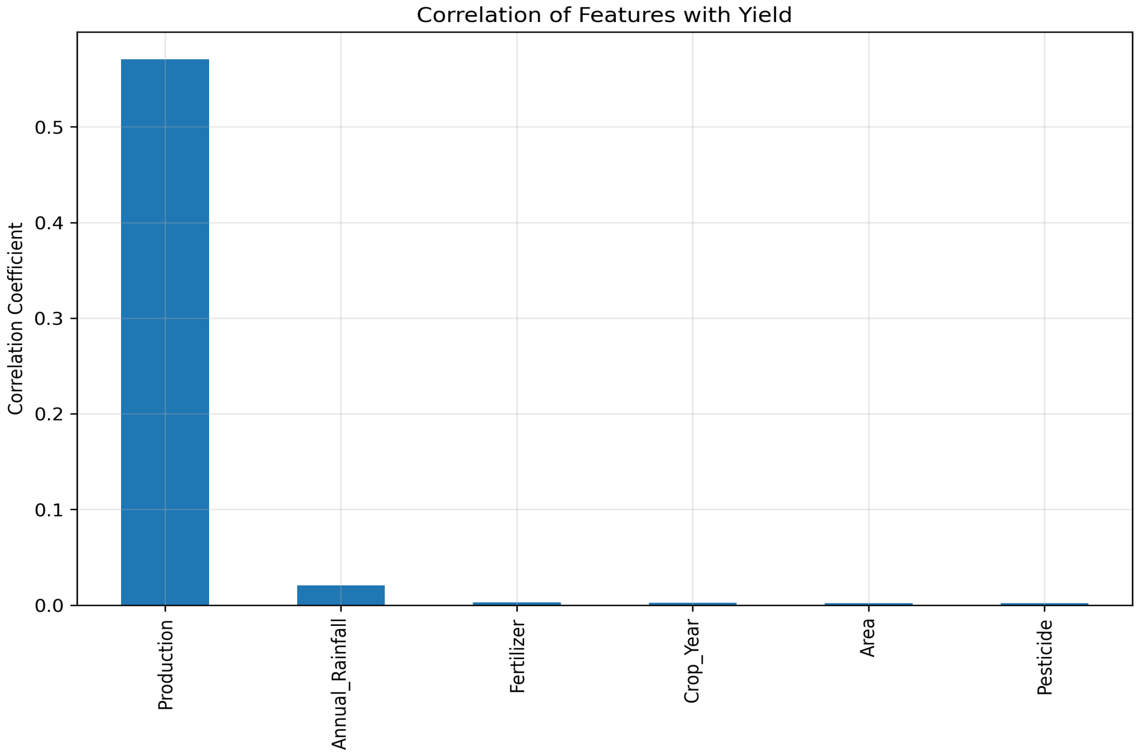

Figure 5: Variables correlated with crop yield

Observations on Correlations: The correlation analysis reveals that Production has the strongest positive correlation with Yield (0.57), which is expected since these variables are directly related. Interestingly, Annual_Rainfall shows only a very slight positive correlation with Yield (0.02), suggesting that while rainfall is important, other factors may play more significant roles in determining crop yield. The correlations between Yield and input factors like Fertilizer (0.0029) and Pesticide (0.0018) are surprisingly weak, indicating that simply increasing these inputs may not necessarily improve yields. This could suggest that optimal application rather than quantity is more important, or that other unmeasured factors (like soil quality or farming techniques) have greater influence. The correlation matrix also shows strong positive correlations between Area and Fertilizer (0.88) and between Fertilizer and Pesticide (0.92), indicating that larger farms tend to use more fertilizer, and farms that use more fertilizer also tend to use more pesticide.

3. Finding Relationships

Further analysis explored the relationships between specific variables and crop yield.

3.1 Impact of Agricultural Inputs on Yield

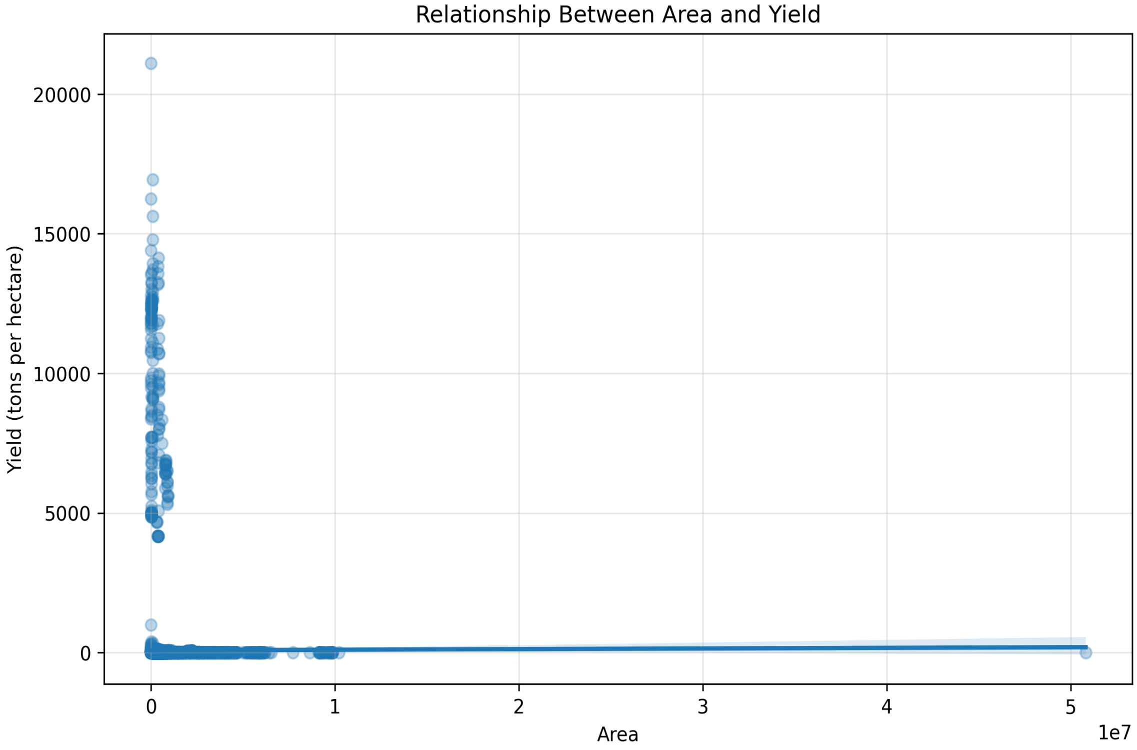

Figure: Relationship between Area and Yield

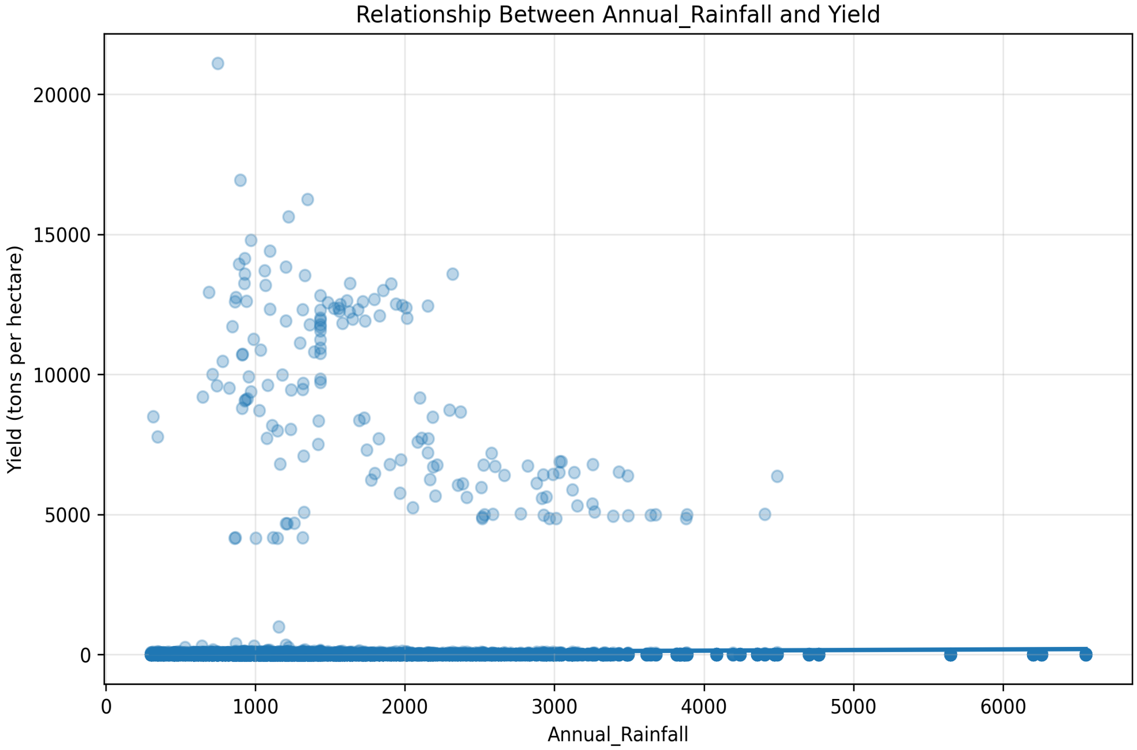

Figure: Relationship between Annual_Rainfall and Yield

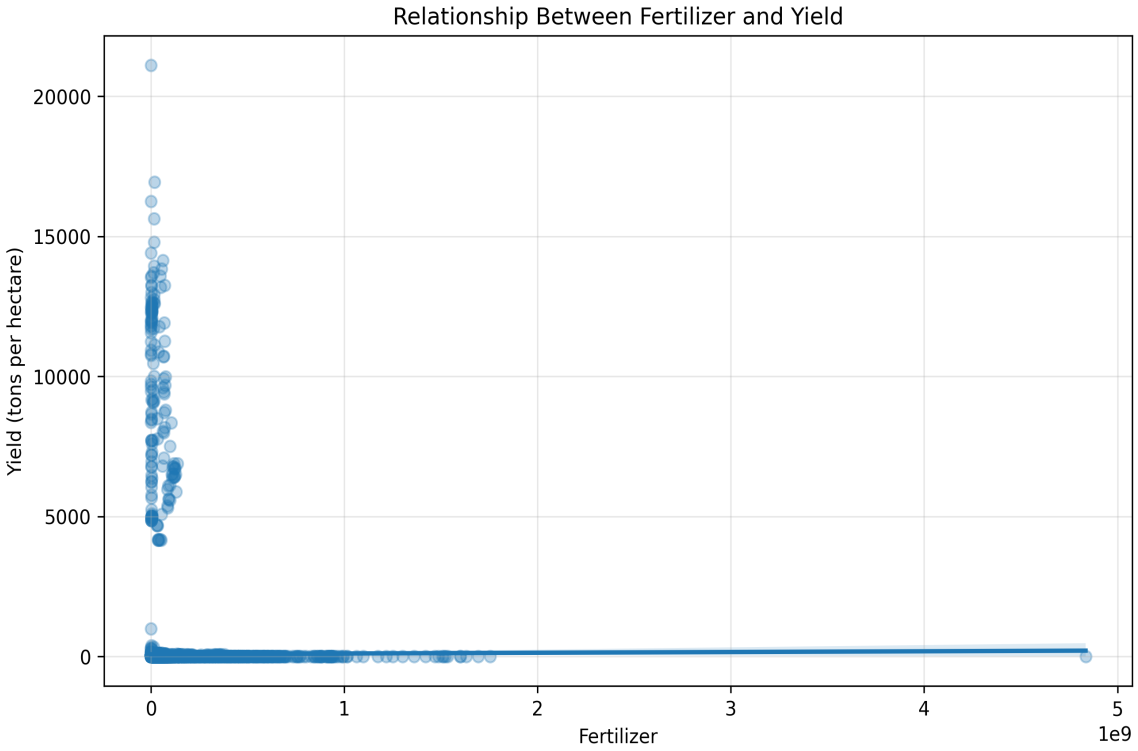

Figure: Relationship between Fertilizer and Yield

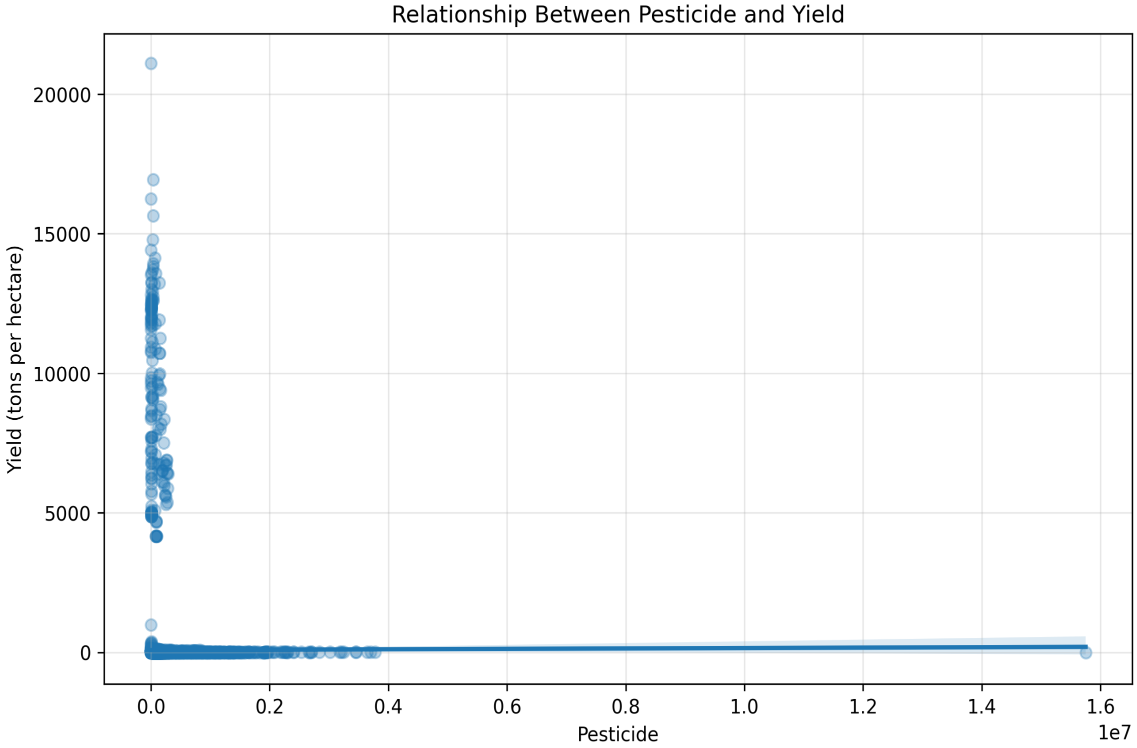

Figure: Relationship between Pesticide and Yield

Observations on Agricultural Inputs: The scatter plots examining relationships between agricultural inputs and yield reveal several important patterns: **Area vs. Yield:** There is no clear linear relationship between farm area and yield. This suggests that farm size alone does not determine productivity, and efficient farming practices can be implemented regardless of farm scale. **Annual Rainfall vs. Yield:** While there is a slight positive trend, the relationship is not strong. This indicates that while adequate rainfall is necessary, excessive rainfall may not proportionally increase yields and may even be detrimental for certain crops. **Fertilizer vs. Yield:** Despite expectations, there is no strong positive correlation between fertilizer application and yield. This counter-intuitive finding suggests that optimal fertilizer usage depends on specific crop requirements, soil conditions, and application methods rather than simply the quantity applied. **Pesticide vs. Yield:** Similar to fertilizer, pesticide usage shows a weak relationship with yield. This may indicate that targeted pest management strategies could be more effective than broad application, or that over-application might even harm beneficial organisms that contribute to crop health.

3.2 Crop Yield by Season

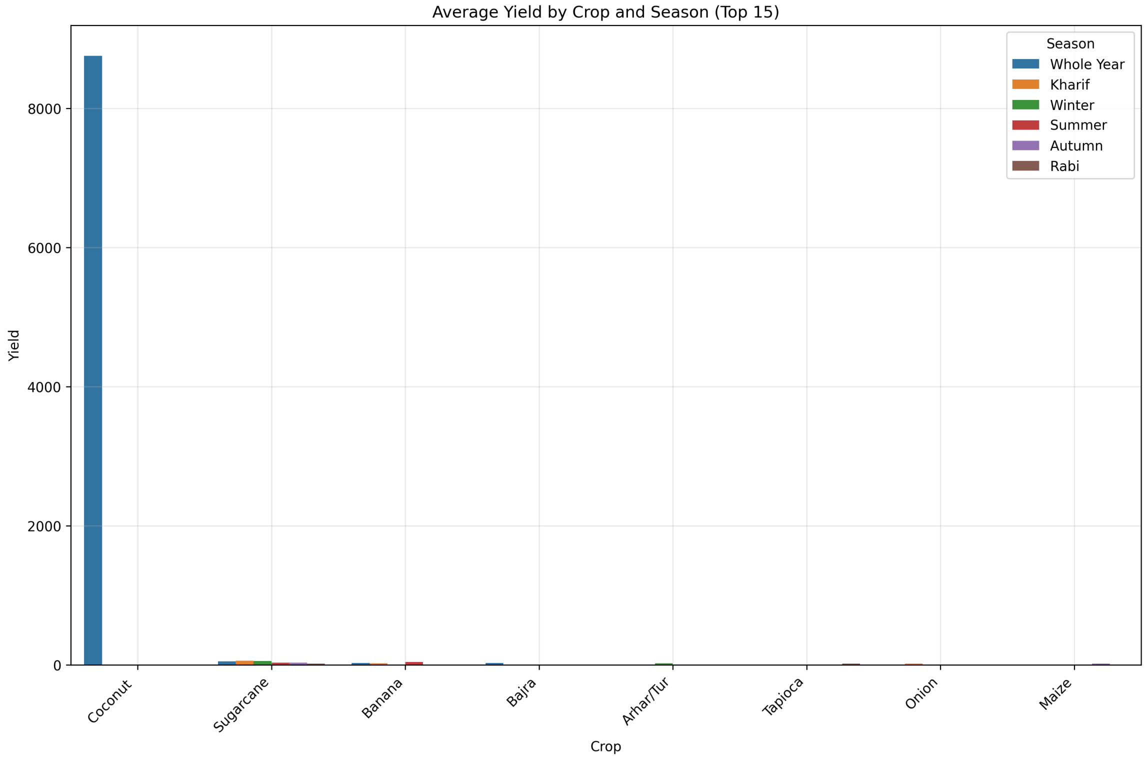

Figure: Average yield by crop and season for top performers

Observations on Crop-Season Combinations: The analysis of yield by crop and season reveals important patterns for maximizing productivity: - Coconut grown throughout the Whole Year consistently shows the highest yields, significantly outperforming all other crop-season combinations. - Sugarcane performs best during the Kharif (monsoon) season, likely due to its high water requirements. - Several crops show distinct seasonal preferences, with some performing significantly better in specific seasons. - The "Whole Year" cultivation approach generally produces higher yields for crops that can be grown in this manner, suggesting advantages to continuous cultivation when possible. These findings highlight the importance of matching crop selection with appropriate growing seasons to maximize yield potential.

4. Clustering Analysis

Farms were clustered based on area, annual rainfall, fertilizer, and pesticide usage to identify groups with similar characteristics and understand patterns in farming practices across India.

4.1 Determining Optimal Number of Clusters

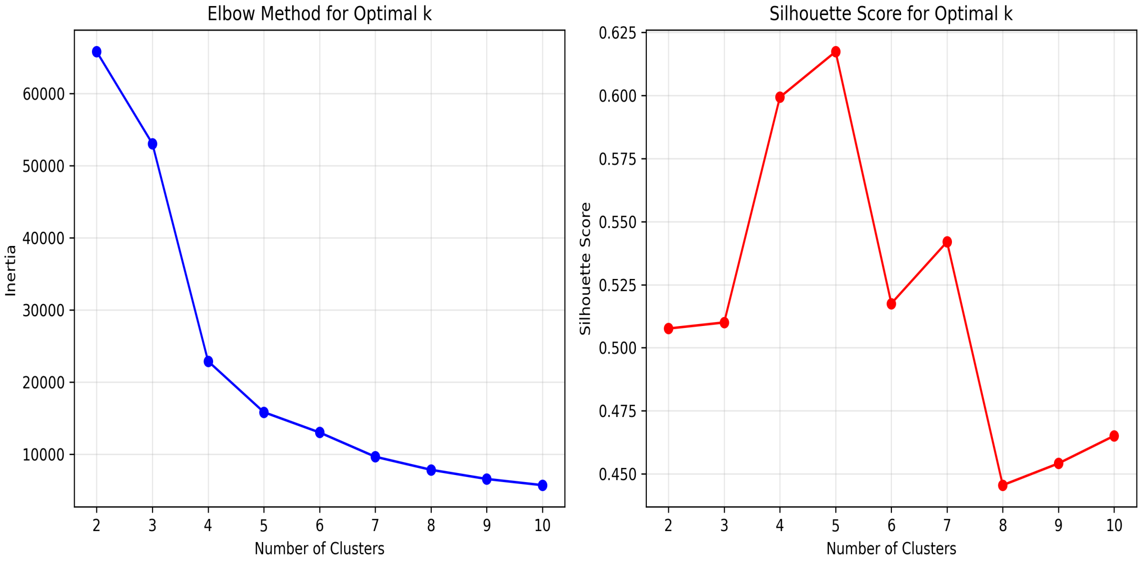

Figure: Elbow method and silhouette scores for determining the optimal number of clusters

Observations on Cluster Evaluation: The cluster evaluation metrics guided the selection of the optimal number of clusters for this analysis. The elbow method plot shows the reduction in inertia (within-cluster sum of squares) as the number of clusters increases. The silhouette score, which measures how well samples fit within their assigned clusters, peaks at 5 clusters, indicating this is the optimal number for segmenting the farms in our dataset. This balance provides meaningful differentiation between farm types while avoiding over-segmentation.

4.2 Cluster Visualization

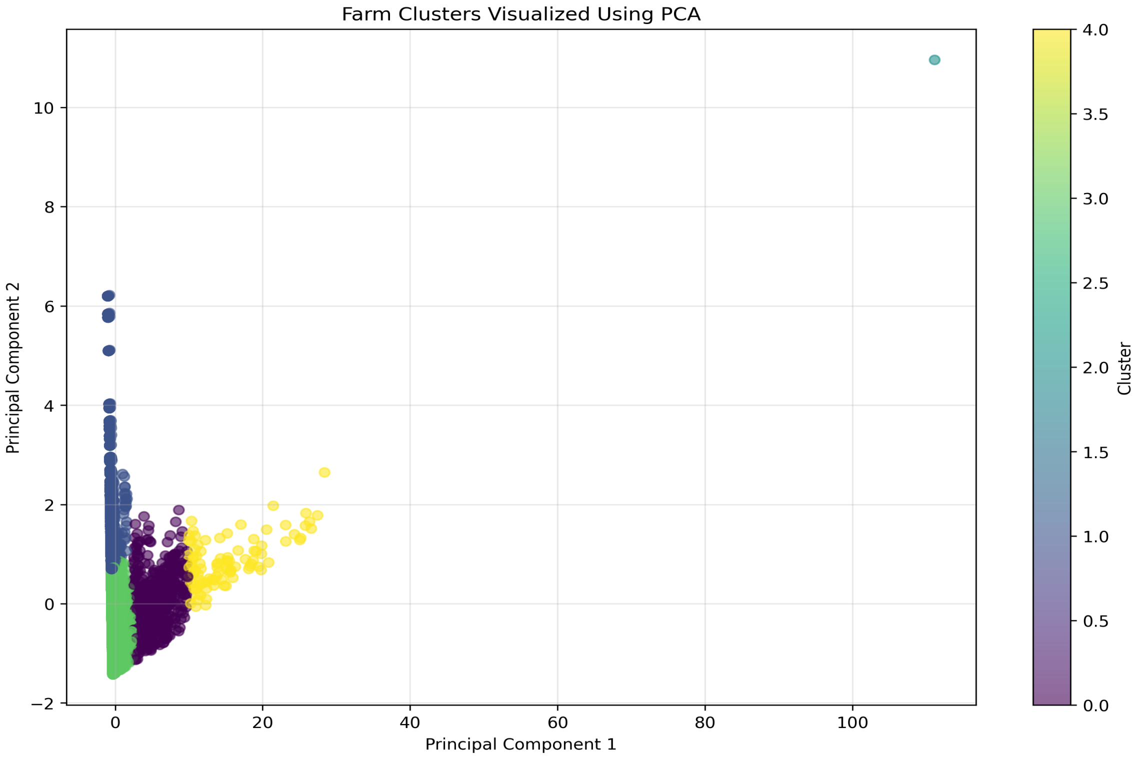

Figure: Visualization of clusters using Principal Component Analysis

Observations on Cluster PCA: The Principal Component Analysis (PCA) visualization reduces the dimensionality of our clustering features to two components, allowing us to visualize the farm segments. The distinct clusters are visible, with Cluster 3 (in green) being the largest and most dominant group, accounting for over 80% of all farms. The smaller clusters are more specialized farm types with distinct characteristics. The PCA plot shows some overlap between clusters, indicating gradual transitions between different farming systems rather than completely discrete categories.

4.3 Cluster Characteristics

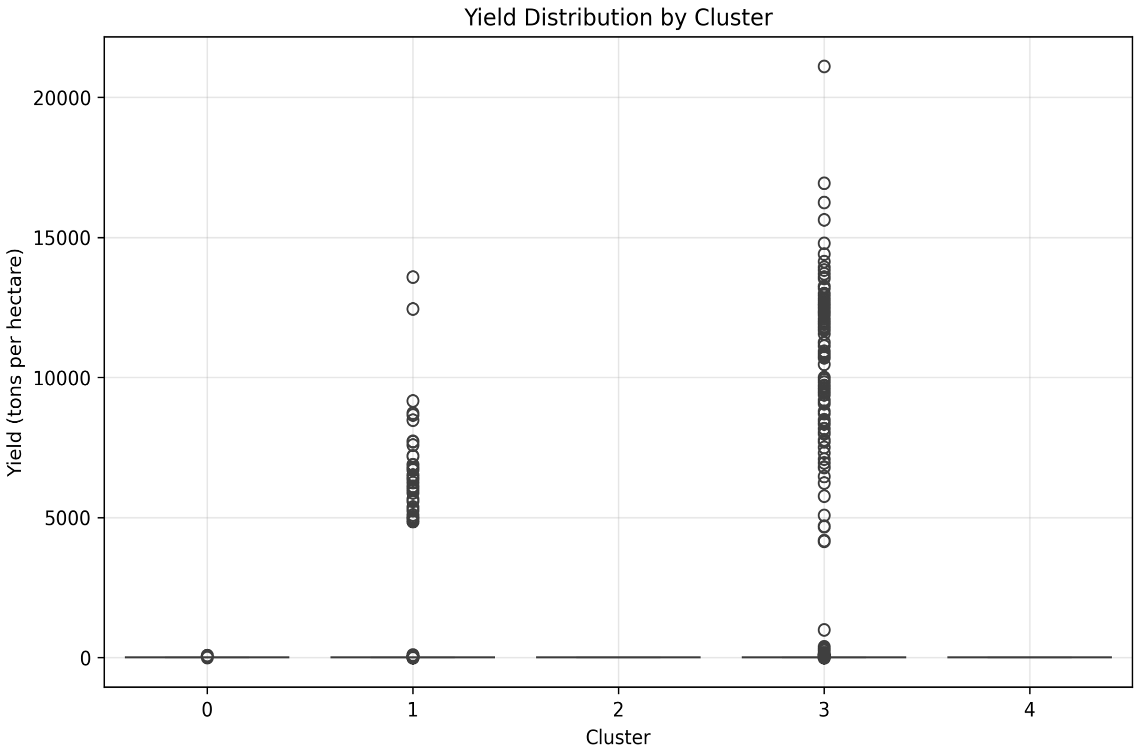

Figure: Yield distribution across different clusters

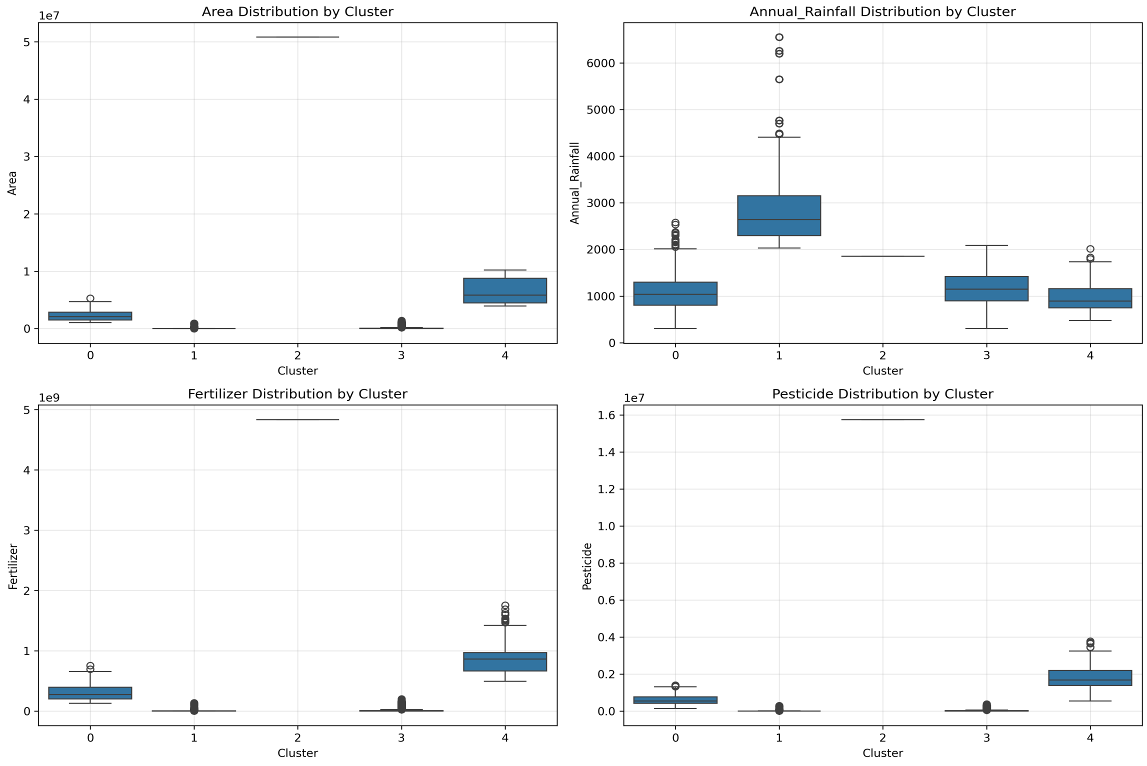

Figure: Distribution of key features across clusters

Observations on Cluster Characteristics: The analysis identified 5 distinct farm clusters with the following key characteristics: **Cluster 0 (610 farms, 3.1%):** These are very large farms with moderate rainfall and extremely high fertilizer usage. With an average yield of just 4.23, these farms appear to be inefficient despite high resource inputs. Sugarcane, Wheat, and Maize are the predominant crops. **Cluster 1 (3200 farms, 16.3%):** This cluster represents medium-sized farms with high rainfall and lower fertilizer usage. They achieve impressive average yields of 115.99, primarily growing Coconut, Sugarcane, and Tapioca. These farms appear to be leveraging natural rainfall effectively to minimize fertilizer inputs while maintaining high productivity. **Cluster 2 (1 farm, 0.005%):** This outlier cluster contains just one extremely large farm with high rainfall and extraordinarily high fertilizer usage. Despite the massive inputs, it achieves a low yield of 0.70, growing Niger seed. This unusual case may represent experimental farming, data error, or very specific conditions. **Cluster 3 (15782 farms, 80.2%):** The dominant cluster consists of smaller farms with moderate rainfall and fertilizer usage. They achieve good average yields of 76.05, primarily growing Coconut, Sugarcane, and Banana. This cluster represents the typical farming system in India. **Cluster 4 (96 farms, 0.5%):** These are large farms with low rainfall and high fertilizer usage. They achieve relatively low yields of 2.15, primarily growing Wheat, Rice, and Cotton. These farms may be struggling with water scarcity despite trying to compensate with higher fertilizer application. The clustering analysis reveals that farm size and input intensity do not necessarily correlate with higher yields. The most successful farms (Cluster 1) achieve high yields with moderate inputs and favorable natural conditions.

5. Recommendations

5.1 Crop Selection

High-yielding crops: Coconut (8652.00), Sugarcane (51.73), and Banana (26.85) consistently show the highest yields across India and should be prioritized where growing conditions permit. **Seasonal considerations:** • Coconut performs best when grown throughout the whole year • Sugarcane shows best results during the Kharif (monsoon) season • Banana yields are highest during the Summer season **Regional preferences:** • Coconut yields are highest in Telangana • Sugarcane performs well in Puducherry • Banana shows exceptional yields in Gujarat

5.2 Optimal Growing Conditions

Coconut: Best with moderate rainfall (around 746 mm annually) **Sugarcane:** Performs well with higher rainfall (around 1,330 mm annually) **Banana:** Thrives with moderate rainfall (around 1,220 mm annually)

5.3 Resource Optimization

Fertilizer usage: Optimize based on crop type. High-yield crops like Coconut require moderate fertilizer application, while some crops may need more intensive fertilization. The weak correlation between fertilizer amount and yield suggests that application method and timing may be more important than quantity. **Rainfall considerations:** Choose water-intensive crops in high-rainfall areas and drought-resistant varieties in low-rainfall regions. The analysis shows that crops have specific rainfall requirements for optimal performance. **Land utilization:** Farm size appears to have minimal correlation with yield, suggesting that efficient farming practices may be more important than farm size. Small-scale farmers can achieve high yields with proper crop selection and resource management.

5.4 Cluster-specific Recommendations

For farms in Cluster 1 (high rainfall, lower fertilizer usage): Ideal for water-intensive crops like Coconut and Sugarcane. Consider increasing fertilizer usage for potentially higher yields, but maintain the current emphasis on leveraging natural rainfall. **For farms in Cluster 3 (moderate rainfall, smaller areas):** Focus on high-value crops like Coconut, Sugarcane, and Banana that perform well in these conditions. This typical farming system in India shows good balance between inputs and outputs. **For farms in Cluster 0 and 4 (lower rainfall, larger areas):** Consider drought-resistant crops and optimize fertilizer usage for cost efficiency. These farms should focus on water conservation techniques and may need to reconsider their crop mix to improve yields.

6. Limitations of Analysis

1. This analysis does not account for soil quality or type, which can significantly affect crop yields. 2. Weather variations beyond annual rainfall (such as temperature, humidity, and seasonal distribution of rainfall) are not considered. 3. Market forces affecting crop selection and economic viability are not included in the analysis. 4. The analysis assumes current agricultural practices and technologies and does not account for adoption of new farming methods. 5. The clustering analysis may be influenced by outliers, particularly in clusters with very few farms. 6. The data spans multiple years but does not specifically analyze year-over-year trends or the impact of climate change.