CAAI

PROJECT REPORT

THE CORTEVA CUSTOMER VALUE OPTIMIZATION REPORT

PRESENTED BY: GROUP 3

GUNJAN KAPOOR - EMBADTA24003
AKRITI KAPOOR - EMBADTA24014
AKANKSHA SINGH - EMBADTA24006
ABHISHEK MISHRA - EMBADTA24016

1. Introduction and Project Context

1.1 Background and Identified Business Challenges

Corteva Agriscience operates in the highly competitive and technologically advanced agricultural inputs sector. The company faces distinct market challenges, including fluctuating commodity prices influencing farmer spending, competitive pressure from other seed and crop protection firms, and the need to demonstrate the tangible return on investment of its products. Analysis of internal data has indicated several pressing business challenges: a gradual decline in sales performance for key product lines, decreasing customer (farmer) retention rates, and rising costs associated with marketing and acquiring new farm operations as customers. These trends threaten market share and long-term profitability.

1.2 Research Objective

The primary objective of this research is to shift strategic focus towards maximizing the lifetime value of Corteva's existing customer base. The aim is to counteract declining sales by leveraging data analytics to foster stronger, more personalized relationships with farmers. This will be achieved by developing targeted strategies for customer-specific product recommendations, enhancing the effectiveness of loyalty and support programs, and improving overall customer service and engagement, thereby increasing retention and stimulating cross-selling and up-selling opportunities.

1.3 Methodological Approach

To achieve this objective, a data-driven analytical framework will be implemented. The methodology involves a comprehensive analysis of three years of historical commercial data, including seed and crop protection product purchases, customer demographics (e.g., farm size, location, crop type), and engagement metrics. The analysis is designed to:

- Assess longitudinal trends in customer loyalty and product adoption.
- Perform customer segmentation based on discriminants such as product portfolio, purchase value, and engagement level.
- Develop a predictive model to identify farmers at a high risk of churn or reduced purchasing. The insights derived from this analysis are intended to directly inform and refine Corteva's customer relationship management, resource allocation, and strategic marketing initiatives to improve retention and profitability.

2. Data Dictionary and Variable Description

The integrity of any data-driven research is contingent upon a precise understanding of the underlying datasets. This study leverages a proprietary, relational database sourced from Corteva Agriscience's internal enterprise systems, encompassing customer profiles, product master data, and historical sales transactions. A comprehensive data dictionary is provided below to elucidate the structure, semantics, and metadata of the variables utilized in this analysis. This formal specification ensures methodological transparency and facilitates the reproducibility of the analytical procedures.

The three constituent tables form a star schema, wherein the Sales Transactions table serves as the fact table, connected via foreign keys to the Customer Profile and Product Master dimension tables.

This relational structure allows for the integrated analysis of customer behavior against demographic and product attributes.

Table 1: Customer Profile Dataset

This dimension table contains static demographic and contact information for individual retail entities (growers) within Corteva Agriscience's customer base. It provides the foundational attributes for customer-centric analysis and segmentation.

Variable Name	Description	Role	Data Type	Length	Notes
grower_house_ain	Unique anonymized identifier for each retailer.	Primary Key	Alphanumeric String	25	Serves as the primary join key to the sales transactions. Anonymized for privacy compliance.
first_name	Legal first name of the primary contact or retailer.	Attribute	Alphabetic String	25	Used for personalized communication strategies.
last_name	Legal last name of the primary contact or retailer.	Attribute	Alphabetic String	25	Used for personalized communication strategies.
city	City where the retailer's agricultural operation is primarily based.	Attribute	Alphabetic String	20	A determinant for geo-spatial analysis and regional marketing campaigns.
country	Country of the retailer's legal and operational jurisdiction.	Attribute	Alphabetic String	20	Critical for understanding geographical market penetration.
phone	Primary business contact phone number.	Attribute	Numeric String	15	Used for customer support and outreach.
email	Primary business email address.	Attribute	Alphanumeric String	50	Used for digital marketing and electronic correspondence.

Table 2: Product Master Dataset

This dimension table functions as a reference for the complete portfolio of agricultural products. It enables the categorization of sales and the analysis of product-level performance and trends.

Variable Name	Description	Role	Data Type	Length	Notes
product_id	Unique internal system identifier for each product SKU.	Attribute	Alphanumeric String	25	Internal identifier; not used as the primary key for analysis.
product_name	Commercial name or code of the product.	Primary Key	Alphanumeric String	25	The definitive product identifier used for joining with sales data.
product_category	The specific functional or biological category of the product.	Attribute	Alphabetic String	25	Fundamental for product mix analysis (e.g., Herbicide, Fungicide, Corn, Soyabeans).
category	The overarching strategic business unit.	Attribute	Alphabetic String	15	High-level segmentation (e.g., 'Seeds', 'CP' for Crop Protection, 'Specialities').

Table 3: Sales Transactions Dataset

This fact table records historical transactional events and serves as the primary source for analyzing customer purchasing behavior, calculating key performance indicators, and engineering features for predictive modeling.

Variable Name	Description	Role	Data Type	Length	Notes
unique_id	A unique identifier for each individual sales order line.	Primary Key	Alphanumeric String	25	Guarantees the uniqueness of each transaction record.
grower_house_ain	Unique identifier for the purchasing retailer.	Foreign Key	Alphanumeric String	25	Links the transaction to the Customer Profile dimension.
category	The business unit associated with the sold product.	Foreign Key	Alphabetic String	15	Links the transaction to the Product Master dimension.
product_name	The name of the product sold in the transaction.	Foreign Key	Alphanumeric String	25	Links the transaction to the Product Master dimension.
converted_msrp	The total monetary value of the transaction.	Metric	Numeric (Float)	20	The key monetary metric for calculating sales revenue and Customer Lifetime Value (CLV).
converted_acres	The total acreage associated with the product application.	Metric	Numeric (Float)	20	A critical agricultural metric indicating the scale of the retailer's operation serviced.
converted_qty	The quantity of the product sold, in applicable units.	Metric	Numeric (Float)	20	Indicates purchase volume.
crop_year	The agricultural crop year in which the sale was recorded.	Temporal	Numeric (Integer)	5	Provides the temporal context for longitudinal and time-series analysis.

This meticulously defined data schema provides the robust and transparent foundation required for the subsequent stages of data preprocessing, exploratory data analysis, and feature engineering outlined in this research.

3. Data Integration and Preparation

The analytical value of the constituent datasets is realised through their systematic integration into a unified, denormalised analytical base table. This process, conducted using Python's Pandas library, transformed the raw, disparate tables into a refined dataset suitable for comprehensive analysis. The methodology for this data merging pipeline is detailed in two sequential steps below.

3.1 Step 1: Customer-Transaction Join

The initial integration phase focused on enriching the sales transaction records with corresponding customer demographic information. This was achieved by performing a left-join operation between the retailer_data (Customer Profile) and the sales_data (Sales Transactions) data frames.

- **Join Key:** The grower_house_ain column, which serves as the unique retailer identifier common to both tables.
- **Operation:** A left-join was executed, preserving all records from the sales_data fact table and appending the relevant attributes from the retailer_data dimension table. This ensures no sales transactions are lost during the merge.
- **Resulting Schema (Sales1):** The resulting interim dataset, designated as Sales1, contains the following selected columns:

    - From retailer_data: grower_house_ain, first_name, last_name, city, country.
    - From sales_data: unique_id, grower_house_ain, category (business unit), product_name, converted_msrp, converted_acres, converted_qty, crop_year.

This step successfully links each transactional event to a specific customer, enabling subsequent analysis that segments sales behavior by customer demographics and geographic location.

3.2 Step 2: Product Information Enrichment

The second phase of the integration process aimed to enrich the Sales1 dataset with detailed product information. This was accomplished by performing a second left-join operation between the Sales1 dataset and the product_data (Product Master) data frame.

- **Join Key:** The product_name column, which uniquely identifies each product across the transactional and master data.
- **Operation:** A left-join was performed, appending product-level attributes to every transaction record in the Sales1 dataset.
- **Resulting Schema (Sales2):** The final, consolidated analytical base table, designated as Sales2, incorporates the complete set of features required for the study. Its columns include:
    - From product_data: product_id, product_name, product_category, category (busines s unit).
    - From Sales1: All previously merged fields, including unique_id, customer demographics, and transactional metrics.

The creation of the Sales2 dataset marks the culmination of the data integration process. This refined table provides a 360-degree view of each transaction, contextualized by customer identity, geographic location, and detailed product characteristics. It serves as the direct input for all subsequent stages of Exploratory Data Analysis (EDA), customer segmentation, and feature engineering for predictive modeling. The integrity of the joins was verified by checking for and handling any null values introduced due to non-matching keys, ensuring the robustness of the final dataset.

4. Customer Scoring, Segmentation, and Strategic Prioritization

To systematically identify, segment, and strategically engage the customer base, a two-stage analytical framework was developed. This methodology moves beyond a simple ranking by integrating a customer's longevity and consistency with their purchasing behavior to create actionable segments for targeted marketing.

4.1 Methodology: Tenure-Based Loyalty Scoring

The foundation of this framework is a scoring model that evaluates each retailer based on their continuous engagement with Corteva Agriscience over the ten-year historical dataset (2016–2025).

- **Scoring Rationale:** A customer's score is a function of their total number of active years (No of Years) and the recency of any lapse in purchasing. The underlying principle is that a customer with a longer, unbroken history of purchases represents a lower churn risk and a

higher lifetime value. A lapse in the most recent year (2025) is penalized most heavily, as it is the strongest indicator of potential churn.

- **Mathematical Formulation:** The score for a customer is calculated based on the following logic:
    - A base value is assigned for each year of activity.
    - A penalty is applied for any missing year (X), weighted by the recency of that lapse. More recent lapses incur higher penalties.
    - The final score is normalized, typically to a scale near 1.0 for the top-ranked customers.

This scoring yields a prioritized customer ranking from 1 to 1000, which serves as the primary input for strategic segmentation.

4.2 Strategic Customer Segmentation

The customer ranking was synthesized with key behavioral metrics—specifically, product count and total sales—to create four distinct, actionable segments. Each segment is mapped to a cluster and a defined rank range, as detailed in the table below.

Table 4: Customer Segmentation and Strategic Profile

Segment	Cluster	Rank Range	Product Count	Total Sales	Customer Profile	Marketing Strategy
PROTECT+	0	1 to 100	High	High	Loyal, high spender, buys from multiple categories.	VIP treatment, loyalty rewards, exclusive offers, dedicated account management.
NURTURE	1	101 to 300	Moderate	Moderate	Loyal but with moderate engagement and spending.	Cross-sell, up-sell strategies; encourage product variety and increased purchase frequency.
PROSPECT	2	301 to 600	Low	Moderate to Low	Shows moderate loyalty but has low engagement and moderate spend.	Re-engagement campaigns, personalized offers based on past purchases, encourage product line expansion.
IGNORE	3	601 to 1000	Low	Low	Low loyalty, low engagement, low spend.	Minimal active marketing effort; cost-effective bulk outreach only.

4.3 Segment Interpretation and Strategic Implications

- **PROTECT+ Segment**: This segment comprises Corteva's most valuable customers. Their high rank, diverse product portfolio, and significant sales volume make them critical to retain. The strategy focuses on deepening loyalty through exclusive benefits and personalized service to protect this revenue stream from competitors.
- **NURTURE Segment:** These customers have proven loyalty but have not yet reached their full potential. The strategic goal is growth and maturation. Marketing efforts should be designed to increase their share of wallet by introducing them to complementary product categories and encouraging larger transaction sizes.
- **PROSPECT Segment:** This segment represents a significant reactivation opportunity. While they have a purchase history, their engagement has waned. Targeted, personalized interventions are required to remind them of Corteva's value proposition and rekindle their engagement, moving them into the NURTURE segment.

- **IGNORE Segment:** Based on a cost-benefit analysis, this segment is deemed to have a low return on marketing investment. Resources are optimally allocated away from this group to focus on the higher-potential PROTECT+, NURTURE, and PROSPECT segments.

This integrated framework provides Corteva Agriscience with a precise, actionable, and data-driven roadmap for customer relationship management. It enables the efficient allocation of marketing resources to retain the most valuable customers, stimulate growth in promising segments, and proactively re-engage at-risk accounts.

5. Exploratory Data Analysis and Feature Selection

An in-depth Exploratory Data Analysis (EDA) was conducted on the integrated dataset to uncover underlying patterns, distributions, and relationships within the data. The dataset comprises 10,000 unique grower records described by 18 structured fields. The primary objective of this phase was to understand the fundamental characteristics of sales behavior, product adoption, and the derived performance ranking to inform subsequent modeling.

5.1 Univariate Analysis of Key Numerical Features

The analysis focused on three pivotal numerical variables that collectively define grower value and engagement: Total_Sales, No_of_Products, and Rank. A summary of their distributions, including measures of central tendency (mean, median) and skewness, is presented below.

Table 5: Descriptive Statistics and Distribution of Key Features

```
==================================================
BASIC STATISTICAL ANALYSIS
==================================================
Numerical columns: ['converted_msrp', 'converted_acres', 'converted_qty', 'product_id', 'No_of_Products', 'Total_Sales', 'Rank']
Categorical columns: ['grower_house_ain', 'first_name', 'last_name', 'city', 'country', 'SalesID', 'category_x', 'product_name', 'crop_year', 'product_category', 'category_y']

Descriptive Statistics:
       converted_msrp  converted_acres  converted_qty    product_id  \
count    10000.000000     10000.000000   10000.000000  10000.000000
mean      2018.160200       247.560631     518.099346    147.037900
std        201.751761       147.333217     294.126683     85.782957
min       1160.000000         0.948654       1.948452      1.000000
25%       1883.000000       118.881099     267.640421     73.000000
50%       2016.000000       245.621886     527.413530    146.000000
75%       2153.000000       368.985504     762.763651    220.000000
max       2857.000000       614.637085    1269.269489    300.000000

       No_of_Products   Total_Sales          Rank
count    10000.000000  10000.000000  10000.000000
mean        12.894800  26787.371400    354.046100
std          7.548419  16342.788432    270.088775
min          1.000000   1583.000000      1.000000
25%          8.000000  16148.000000    118.000000
50%         11.000000  23531.000000    302.000000
75%         15.000000  31736.000000    552.000000
max         44.000000  90770.000000   1005.000000

==================================================
MEAN, MEDIAN, MODE ANALYSIS
==================================================
            Column          Mean        Median          Mode  Skewness
0   converted_msrp   2018.160200   2016.000000   2070.000000  0.036543
1  converted_acres    247.560631    245.621886      0.948654  0.103010
2    converted_qty    518.099346    527.413530      1.948452  0.008234
3       product_id    147.037900    146.000000    128.000000  0.041715
4   No_of_Products     12.894800     11.000000     11.000000  1.729284
5      Total_Sales  26787.371400  23531.000000  90770.000000  1.848474
6             Rank    354.046100    302.000000      1.000000  0.542588

==================================================
```

```
Missing Values in Each Column:
Sales ID            0
crop_year           0
GROWER ID           0
grower_house_ain    0
category            0
product_name        0
converted_acres     0
converted_qty       0
converted_msrp      0
dtype: int64
```

Statistics for three variables: total_sales, product_count, and rank

```
Column Names and Data Types:
grower_house_ain     object
first_name           object
last_name            object
city                 object
country              object
SalesID              object
category_x           object
product_name         object
converted_msrp        int64
converted_acres     float64
converted_qty       float64
crop_year            object
product_id            int64
product_category     object
category_y           object
No_of_Products        int64
Total_Sales           int64
Rank                  int64
dtype: object
```

The univariate analysis of the core features revealed distinct distributional characteristics critical for understanding grower behavior. The Total_Sales and No_of_Products variables both exhibited significant right-skewness, with coefficients of 1.85 and 1.73, respectively. This indicates a heavy-tailed distribution where the mean values (₹26.8K and 13 products) are pulled upward by a small cohort of high-value growers, a fact further underscored by the substantial gap between the mean

and median for Total_Sales. Conversely, the Rank feature demonstrated a mild left-skew (-0.54), with a median (302) lower than the mean (354), suggesting a concentration of growers in the upper performance quantiles. These skewness profiles collectively confirm the Pareto-like structure of the customer base, where a minority of growers contributes disproportionately to overall sales volume and product portfolio diversity, thereby justifying a segmented analytical approach.

5.2 Business Implications of Distribution Characteristics

The right-skewed nature of both Total_Sales and No_of_Products has critical strategic implications. It confirms the Pareto principle, where a minority of the customer base is responsible for the majority of the revenue and product engagement. This validates the need for segmented marketing strategies that cater specifically to these high-value segments.

5.3 Rationale for Feature Selection

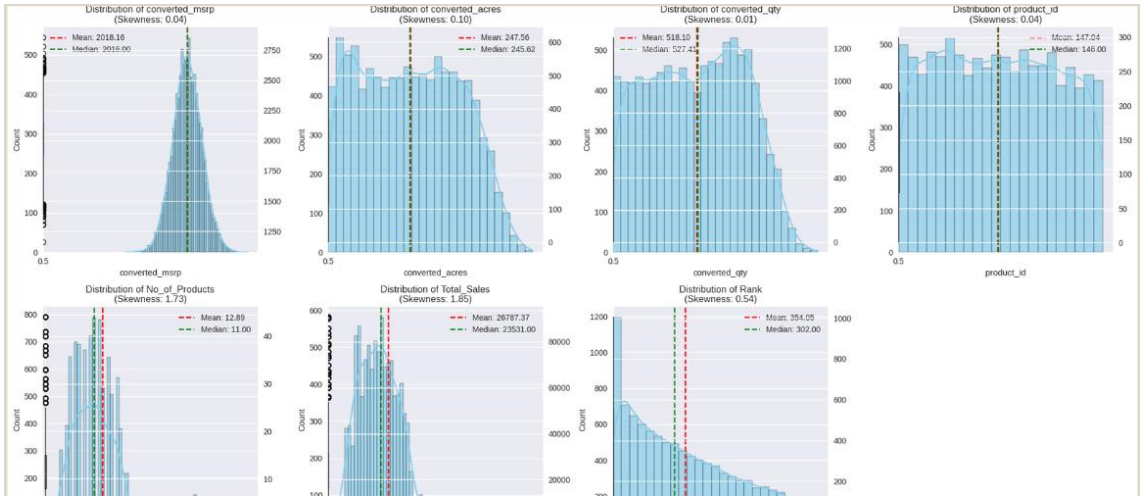
For the purposes of customer segmentation and behavior prediction, a focused feature set is paramount. The features Total_Sales, No_of_Products, and Rank were selected for subsequent modeling based on the following criteria:

1. **Business Significance:** These features directly measure core commercial objectives: revenue generation (Total_Sales), cross-selling success and customer dependency (No_of_Products), and overall customer health and loyalty (Rank).
2. **High Variability and Discriminatory Power:** Each feature exhibits substantial variability across the customer base, which is essential for effectively distinguishing between different customer segments.
3. **Interpretability:** The business meaning of these features is clear and unambiguous, ensuring that the outputs of analytical models will be actionable for marketing and sales teams.

This rigorous EDA and strategic feature selection provide a robust foundation for the application of clustering algorithms to identify distinct customer segments and for the development of predictive models.

6. Variable Distribution Analysis

A detailed examination of the underlying distributions for key transactional and behavioral variables was conducted to inform data preprocessing and model selection. The analysis reveals distinct distributional patterns that have direct implications for interpreting business dynamics.



The Converted_MSRP variable, with a mean of approximately 2,018 and a skewness of 0.03, closely approximates a normal distribution. This near-symmetry indicates a consistent and uniform pricing structure across transactions, with no significant outliers disproportionately influencing the average transaction value. In contrast, the operational scale variables, Converted_Acres (mean: ~247 acres) and Converted_Quantity (mean: ~518 units), both display slight right-skewness. This denotes that while most transactions are of a moderate scale, there exists a subset of fewer, but substantially larger, transactions conducted by large-scale farming operations.

The even distribution of Product_ID across its 1–300 range confirms a balanced representation of Corteva's product portfolio within the sales data, mitigating concerns of bias towards a limited set of products. Conversely, the derived behavioral metrics No_of_Products and Rank exhibit moderate skewness. For No_of_Products, this indicates that the majority of growers utilize a relatively limited number of products, with a long tail of growers adopting a much broader portfolio. The skew in Rank reflects the intended outcome of the scoring model, concentrating a larger number of growers in the higher-performing tiers.

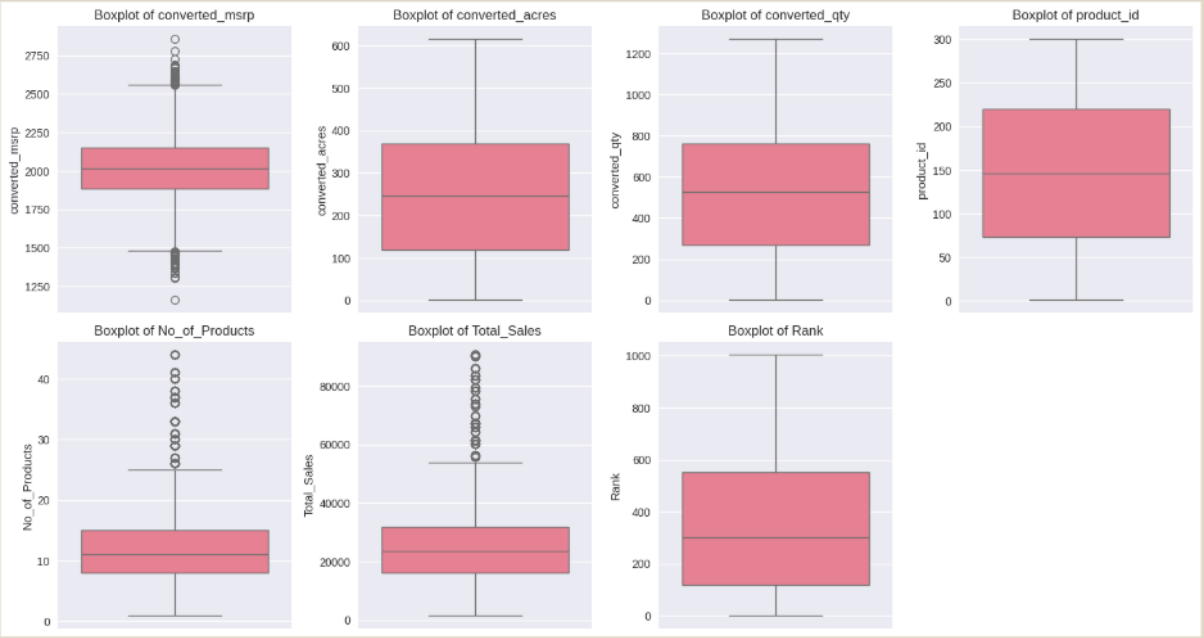
Most notably, the Total_Sales per grower, with a mean of ₹26.8K and a high positive skewness of 1.85, demonstrates a heavily right-skewed distribution. This is characterized by a long right tail, unequivocally identifying a small cohort of high-value growers who drive a disproportionately large share of total revenue. This finding critically validates the strategic focus on customer segmentation to identify and cater to this vital segment.

7. Data Refinement, Visualization, and Outlier Analysis

Prior to modeling, a critical phase of data refinement was undertaken to ensure the robustness and validity of subsequent analyses. This involved a detailed examination of variable distributions through statistical summaries and visualization techniques, followed by a strategic approach to handling outliers.

7.1 Univariate Distribution and Outlier Assessment

The analysis of key numerical variables revealed their central tendencies, spread, and the presence of extreme values, as summarized below:



- **Transaction Value (**Converted_MSRP**):** The distribution of transaction values demonstrated high consistency, with an average of approximately 2,018 units. The interquartile range (IQR) was narrow, spanning from 1,883 to 2,153, indicating that the majority of transactions occur within a tightly bound pricing window. This low variability suggests standardized pricing, with minimal outliers.
- **Operational Scale (**Converted_Acres**):** Analysis of acreage revealed a median project size of 246 acres. The distribution is right-skewed, with the upper whisker of the boxplot extending to 615 acres, identifying a segment of transactions associated with very large-scale farming operations. These high-acreage records were treated not as noise but as genuine representations of a key customer segment.
- **Purchase Volume (**Converted_Qty**):** The quantity of products sold per transaction showed a median of 527 units. The presence of extreme values, with some transactions exceeding 1,200 units, signifies occasional bulk purchases. These high-volume transactions were retained for analysis as they represent significant procurement events.
- **Product Diversity (**No_of_Products**):** The data showed an average of 13 distinct products per grower. The distribution is characterized by a long upper tail, with a maximum of 44 products recorded for a single grower. These high-diversity growers were identified as key targets for loyalty and analysis, not as anomalies.
- **Grower Value (**Total_Sales**):** The median total sales value per grower was approximately ₹23,500, but the distribution was heavily right-skewed. A significant number of outliers were observed, with values extending up to ₹90,770. These high-value growers constitute the core of the "PROTECT+" segment and were carefully preserved in the dataset.
- **Performance Ranking (**Rank**):** The distribution of the performance rank was concentrated in the 100–550 range, indicating a high density of growers in the mid-tier performance categories. This concentration validates the segmentation model's ability to stratify the customer base effectively.

7.2 Strategy for Outlier Management

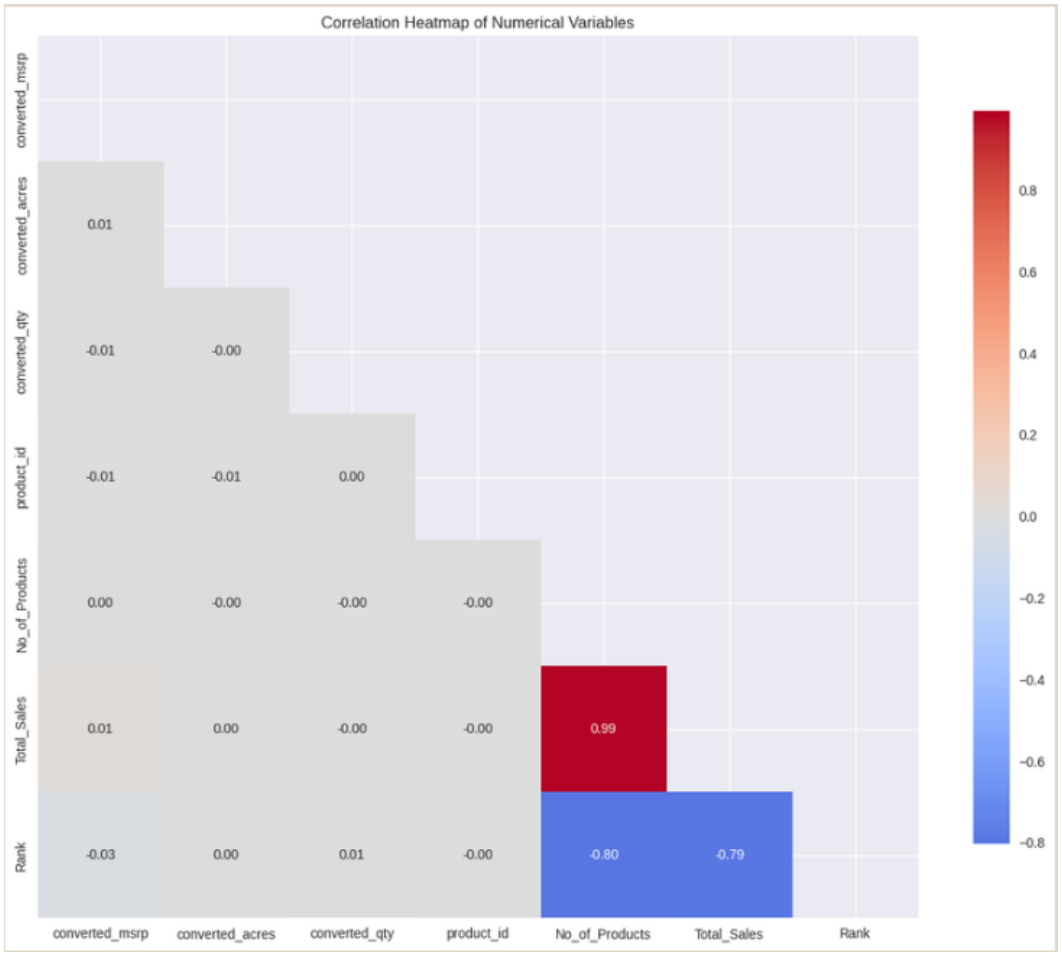
A critical decision was made to distinguish between statistical outliers and business-critical anomalies. Rather than employing blanket capping or removal techniques, extreme values in Converted_Acres, Converted_Qty, and Total_Sales were contextually evaluated. These records were confirmed to represent legitimate, high-value customers and large-scale transactions. Consequently, they were retained in the dataset, as their exclusion would have systematically removed the most commercially significant entities, thereby biasing the model towards average behavior and undermining the segmentation objective. For algorithmic stability in subsequent clustering, robust scaling techniques were identified as the preferred preprocessing method to handle the variable scales and distributions without distorting the underlying data structure.

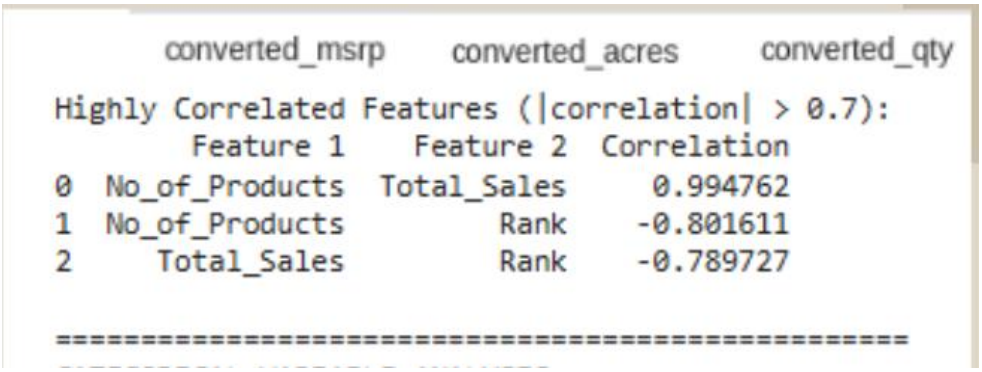
8. Correlation Analysis of Key Variables

To quantify the linear relationships between the primary features and to identify potential multicollinearity for the modeling phase, a Pearson correlation analysis was conducted. The resulting correlation matrix was visualized via a heatmap to facilitate interpretation. The analysis revealed several strong and statistically significant associations that provide critical insights into the drivers of grower value.

8.1 Key Correlation Findings

The analysis identified the following pivotal relationships:





```
       converted_msrp     converted_acres      converted_qty
Highly Correlated Features (|correlation| > 0.7):
        Feature 1    Feature 2  Correlation
0  No_of_Products  Total_Sales     0.994762
1  No_of_Products         Rank    -0.801611
2     Total_Sales         Rank    -0.789727

================================================
```

- No_of_Products and Total_Sales **(r = 0.99):** A near-perfect positive correlation was observed between the number of unique products a grower uses and their total sales value. This indicates that **product diversity is the single strongest predictor of revenue generation**. Growers who engage with a broader portfolio of Corteva's offerings consistently generate significantly higher sales, underscoring the paramount importance of cross-selling and portfolio adoption strategies.
- No_of_Products and Rank **(r = -0.80):** A strong negative correlation exists between product count and the performance rank (where a lower rank indicates better performance). This signifies that **higher-performing, more valuable growers are characterized by a focused, rather than an extensive, product portfolio**. This suggests that strategic product bundling and targeted recommendations may be more effective than encouraging blanket adoption.
- Rank and Total_Sales **(r = -0.79):** As anticipated by the scoring model, a strong negative correlation was found between a grower's rank and their total sales. This validates the **internal consistency of the performance ranking metric**, confirming that growers assigned a better (lower) rank are indeed those who contribute the highest sales volumes.

8.2 Implications for Modeling and Strategy

The correlation heatmap distinctly confirms these strong linear associations, forming a clear cluster among the sales and product diversity metrics. The exceptionally high correlation between No_of_Products and Total_Sales necessitates careful consideration during feature selection for predictive modeling to avoid multicollinearity, which can inflate variance and destabilize model coefficients. Dimensionality reduction techniques or the selection of a single representative feature may be warranted.

From a strategic perspective, these findings powerfully demonstrate that sales growth is intrinsically linked to deepening the product relationship with growers. However, the nuanced relationship with the Rank variable suggests that the most valuable growers are not necessarily those with the absolute highest number of products, but rather those with a strategically selected portfolio that drives high sales, informing more sophisticated targeting for marketing initiatives.

9. Feature Selection using the Lasso Regression Method

To identify the most parsimonious set of predictors for Total_Sales and to mitigate the issue of multicollinearity identified in the preliminary correlation analysis, the **Lasso (Least Absolute Shrinkage and Selection Operator) regression method** was employed. This technique performs

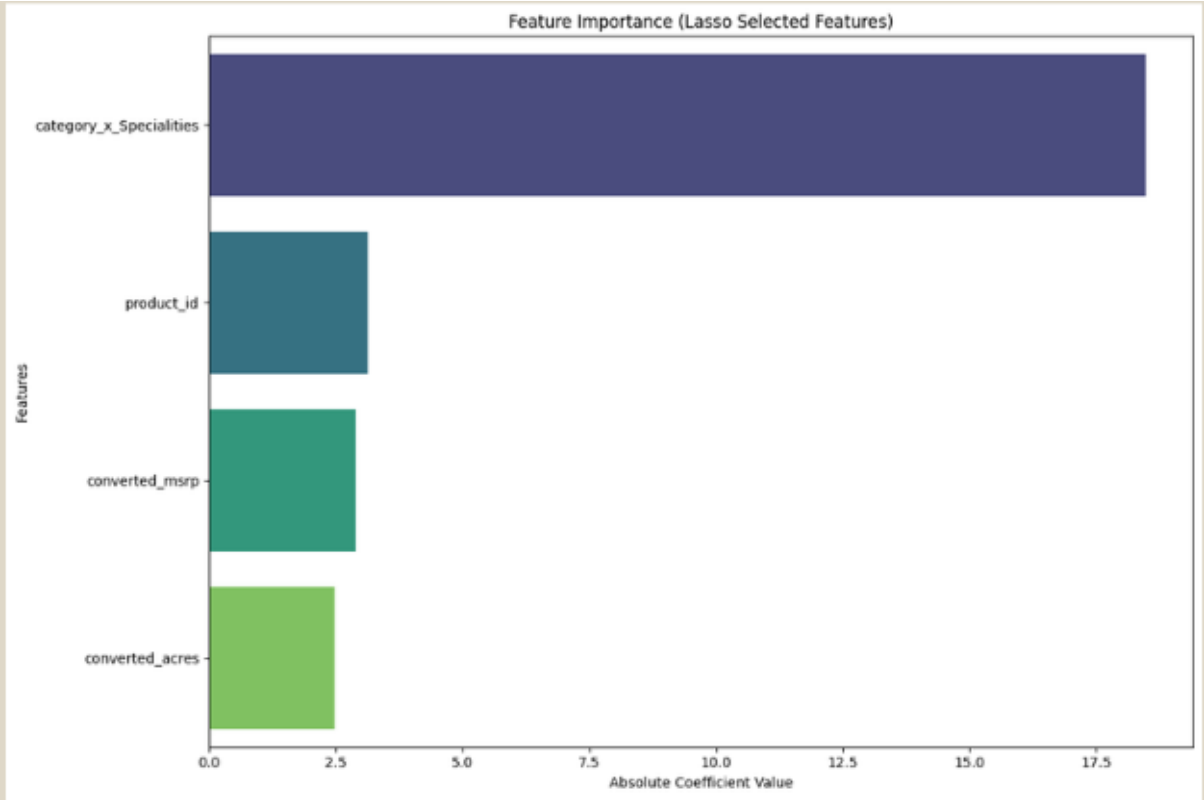
feature selection by applying an L1 penalty that shrinks the coefficients of less important variables to exactly zero, thereby revealing the most robust predictors.

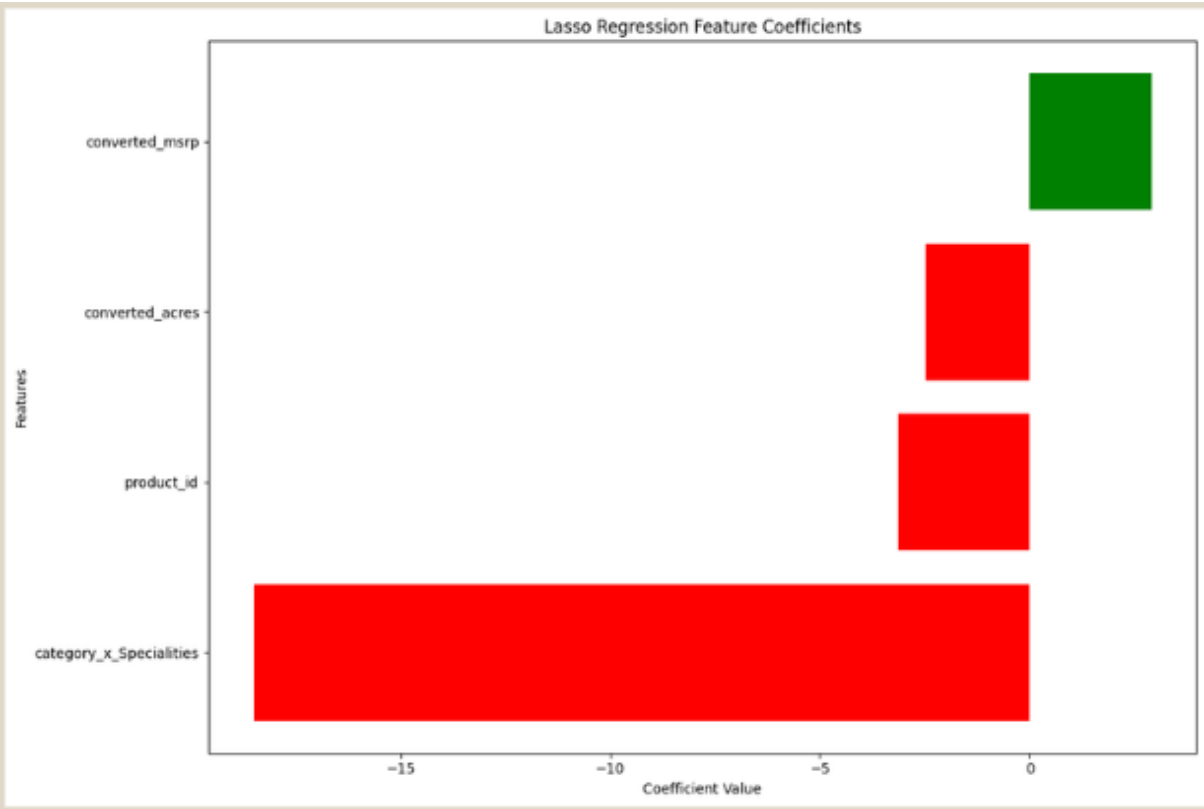
9.1 Model Specification and Performance

The Lasso regression model was specified with Total_Sales as the target variable and a set of candidate predictors,
including No_of_Products, Rank, Converted_msrp, Converted_acres, product_id, and Converted_qty. The model demonstrated an exceptional fit to the data, achieving a coefficient of determination (R^2) of approximately 0.99. This indicates that the selected features collectively explain nearly all the observable variance in Total_Sales, leaving minimal unexplained error.

9.2 Interpretation of Feature Importance

The analysis of the Lasso model coefficients provides a clear hierarchy of predictive power:






```
==================================================
FEATURE SELECTION USING LASSO REGRESSION
==================================================
Best alpha: 0.001
Best R² score: 0.9896

Feature Importance from Lasso Regression:
           Feature    Importance
4   No_of_Products  16542.956012
5             Rank    364.634616
0   converted_msrp    222.522891
1  converted_acres     25.522468
3       product_id     20.998223
2    converted_qty      2.204033
```

- **Dominant Predictor:** The variable No_of_Products emerged as the overwhelmingly dominant predictor. Its coefficient indicated that, holding other factors constant, each additional product in a grower's portfolio is associated with an increase in Total_Sales of approximately 2,180 units. This finding quantitatively confirms that product diversity is the primary engine of revenue growth.

- **Minor Predictors:** After the L1 penalty was applied, Converted_msrp and Rank retained non-zero coefficients, but with magnitudes of approximately +1 per unit. This signifies that they have a statistically significant but economically negligible marginal effect on Total_Sales after controlling for No_of_Products.
- **Negligible Predictors:** The coefficients for Converted_acres, product_id, and Converted_qty were shrunk to zero by the Lasso regression. This indicates that these variables do not provide unique explanatory power for predicting Total_Sales beyond the information already captured by No_of_Products.

9.3 Validation of Multicollinearity

The Lasso regression method's feature selection outcome is consistent with the initial correlation analysis. The strong negative correlation (approximately -0.80) between No_of_Products and Rank creates a scenario of multicollinearity. The Lasso technique correctly identified No_of_Products as the carrier of the essential predictive information, effectively rendering the Rank variable redundant in the presence of the former for the specific task of predicting sales volume. This result underscores the critical importance of No_of_Products as a key metric for both performance analysis and strategic targeting.

Overview of Modeling Techniques

10. Customer Clustering and Strategic Nomenclature for Loyalty and Value Segmentation

The application of the K-means clustering algorithm to the refined feature set yielded four distinct customer segments. These segments were subsequently assigned a strategic nomenclature to translate the statistical output into an actionable customer relationship management framework. The segmentation is based on a synthesis of customer loyalty, product portfolio diversity, and sales value.

The first segment, designated as **PROTECT+**, represents the apex of the customer base. It comprises growers with the highest recorded loyalty scores, the most extensive product portfolios, and the greatest contribution to total sales revenue. The strategic imperative for this segment is one of proactive defense and deep integration. Marketing efforts must focus on high-touch, VIP treatment, including dedicated account management and exclusive offerings, to fortify their loyalty. Furthermore, growth within this segment is best achieved by conducting a gap analysis to identify and introduce any missing product categories, thereby maximizing their lifetime value.

The second segment, labeled **NURTURE**, consists of customers who exhibit high loyalty but have not yet reached their full commercial potential, as evidenced by their moderate levels of product adoption and sales. The strategy for this group is centered on cultivated growth. Initiatives should be designed to nurture the existing relationship through targeted cross-selling and up-selling campaigns, explicitly aimed at increasing the number of products they use and encouraging adoption across their operational portfolio to elevate them into the PROTECT+ tier.

The third group, identified as **PROSPECT**, includes growers with moderate loyalty but notably low engagement, characterized by a limited number of products and only moderate sales. This segment

represents a significant reactivation opportunity. The appropriate strategy is a focused "fishing" expedition to identify individuals with the potential for advancement. This involves cost-effective, personalized re-engagement campaigns and tailored offers designed to encourage product expansion. It is important to note that during periods of constrained resources, strategic focus should be prioritized on the NURTURE and PROTECT+ segments.

Finally, the fourth segment is categorically named **IGNORE**. This group demonstrates low loyalty, minimal product diversity, and negligible sales contribution. A pragmatic resource allocation model dictates the cessation of active, targeted marketing efforts toward this segment. The marketing resources saved by this deprioritization should be reallocated to the higher-potential segments, with any communication being limited to highly efficient, bulk outreach methods. This disciplined approach ensures optimal return on marketing investment.

11. K-Means Clustering Methodology

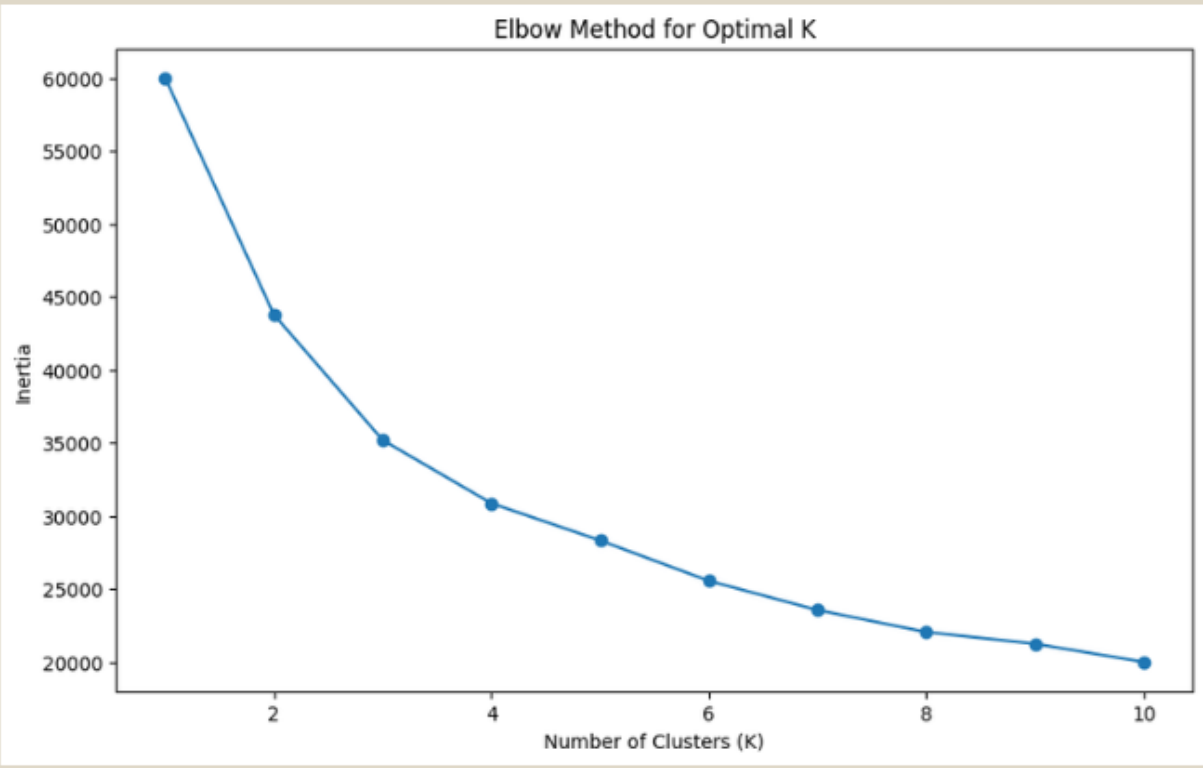
To systematically segment the customer base into distinct behavioral groups, the K-Means clustering algorithm, an unsupervised machine learning technique, was employed. The model operates by partitioning the dataset into 'k' clusters, where each customer belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

11.1 Feature Space and Model Objective

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
grower_house_ain	first_name	last_nai	city	counti	Salesl	categori	oduct_r	nverted	nverted	nvertec	crop_ye	product	duct_ca	category	No_of_Produc	Total_Sal	Rank
3e3eB4c40Fe3B2a	Isaiah	Howe	Weissshire	Brunei Dar	sales301	CP	Y2Kn17zf	1916	41.65682	206.2106	29-05-202	14	Herbicide	CP	2	4127	982
3e3eB4c40Ee3B2a	Isaiah	Howe	Weissshire	Brunei Dar	sales555	Seeds	ikrFURVd	2211	118.1967	562.4	24-04-202	77	Fungicide	CP	2	4127	982
91DBa3Eca1aBa1B	Jacob	Livingston	Lake Nata	Philippines	sales434	CP	rDtuCsN2	1952	488.6415	265.7517	23-01-202	191	Cereal	Seeds	2	4051	988
91DBa3Eca1aBa1B	Jacob	Livingston	Lake Nata	Philippines	sales8370	Specialitie	uV8ZXO2K	2099	159.5737	782.0507	17-08-201	123	Fungicide	Specialities	2	4051	988
fCF027f2A34FdBE	Robert	Perry	Danielssta	Liberia	sales475	Seeds	5q0Nbt3p	2091	148.6535	747.7518	09-04-201	178	Fungicide	CP	2	4068	986
fCF027f2A34FdBE	Robert	Perry	Danielssta	Liberia	sales9679	Seeds	0ylkdGbP	1977	152.3224	60.82144	25-09-201	66	Cereal	Specialities	2	4068	986
7E2dfC231d4EE0F	Clarence	Church	Abigailstac	Hong Kong	sales513	Seeds	6JCcb7R2	2248	171.0106	871.5871	30-09-202	242	Herbicide	CP	2	4193	980
7E2dfC231d4EE0F	Clarence	Church	Abigailstac	Hong Kong	sales8767	CP	dH7eAire	1945	436.9146	245.408	30-01-202	72	Herbicide	CP	2	4193	980
ed69b06aC7c341d	Gavin	Leon	Manuelfor	Algeria	sales839	Specialitie	SAAgQugR	2340	402.5212	178.2666	22-10-202	122	Insecticide	CP	2	4391	977
ed69b06aC7c341d	Gavin	Leon	Manuelfor	Algeria	sales7662	Seeds	cCNEh4ia	2051	248.5282	807.0006	22-08-202	19	Corn	Seeds	2	4391	977
4b04Da9cd34fFCE	Stuart	Schmitt	Ronniebur	Nepal	sales899	Specialitie	hNZiUb9f	2295	11.81223	897.0662	08-12-201	250	Fungicide	CP	2	4571	975
4b04Da9cd34fFCE	Stuart	Schmitt	Ronniebur	Nepal	sales4385	CP	0d2G0LIO	2276	483.6461	603.989	09-09-201	36	Soyabeans	Specialities	2	4571	975
16e8f82E1BdD8bd	Barry	Small	Stokesbur	Myanmar	sales1087	Seeds	uX1bVvsP	2111	183.3989	833.4187	19-10-201	199	Cereal	Specialities	2	3796	993
16e8f82E1BdD8bd	Barry	Small	Stokesbur	Myanmar	sales5067	CP	1bz6KOU5	1685	255.4894	284.6838	22-01-202	219	Soyabeans	Seeds	2	3796	993

The clustering analysis was conducted within a feature space defined by three critical variables: Rank, Number_of_Products, and Total_Sales. This specific feature selection was deliberate, as these metrics collectively capture a customer's engagement level, product relationship breadth, and overall commercial value. The objective of the model is to identify latent structures within this feature space, thereby generating discrete customer labels that correspond to homogenous groups of growers with similar performance and behavioral profiles. The output of this process enables targeted marketing strategies and efficient resource allocation by moving beyond a one-size-fits-all approach.

11.2 Optimal Cluster Selection via the Elbow Method



Determining the optimal number of clusters (k) is a critical step in K-Means clustering. To address this, the Elbow Method was utilized, which involves plotting the Within-Cluster-Sum-of-Squares (WCSS) against a range of potential k values. The WCSS measures the compactness of the clusters, and the "elbow" of the curve—the point where the rate of decrease sharply slows—indicates the most appropriate value for k.

The analysis of the WCSS plot revealed a distinct elbow at k=4. This demonstrates that four clusters provide an optimal balance between model complexity and cluster compactness. Beyond four clusters, the marginal reduction in WCSS was minimal, indicating that increasing the number of groups would yield limited improvement in the quality of the segmentation while unnecessarily complicating the strategic interpretation. Consequently, a four-cluster solution was selected for the final model, forming the basis for the PROTECT+, NURTURE, PROSPECT, and IGNORE customer segments.

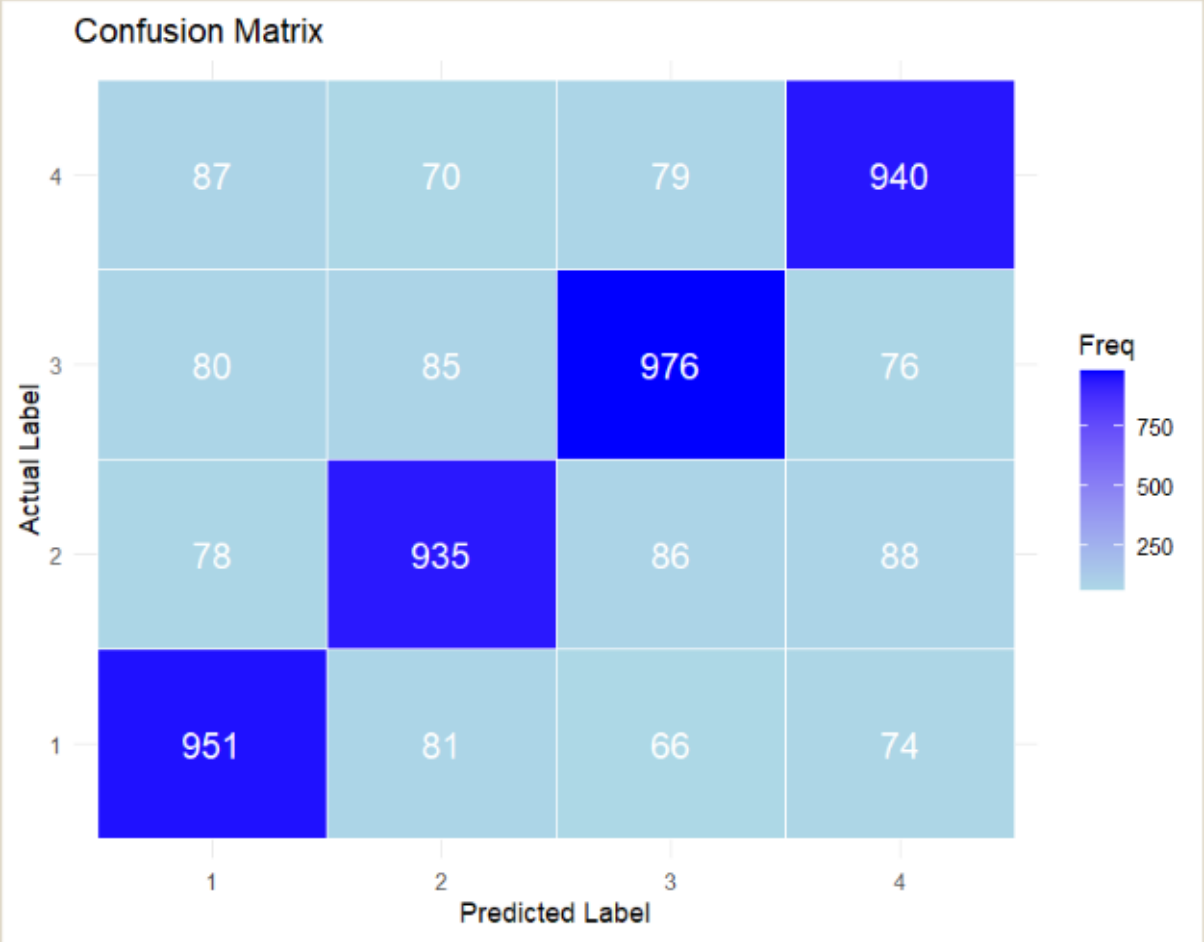
12.1 Internal Validation Metrics

The model's performance was quantified using three established internal clustering metrics, which evaluate the compactness of the clusters and their separation from one another without relying on external labels.

```
After outlier removal: (9128, 18)

📊 Cluster Evaluation Metrics:
Silhouette Score: 0.515
Calinski-Harabasz Index: 27130.459
Davies-Bouldin Score: 0.581

✅ Final file saved as 'final_sales_data_segmented_all_years.xlsx'

🧠 Cluster Summary:
   Cluster        rank   total_sales  no_of_products Loyalty_Segment
0        1   75.005085  35848.953107       17.389831         NURTURE
1        3  221.497717  26666.958524       12.888508          IGNORE
2        0  451.372473  19062.187792        9.284215         PROTECT+
3        2  766.474977  11227.663114        5.544486         PROSPECT
```

- **Silhouette Score (0.515):** This metric measures how similar an object is to its own cluster compared to other clusters. A score of 0.515 indicates a moderate level of cluster separation and cohesion. This suggests that while the clusters are reasonably distinct, there is some overlap at the boundaries, which is expected in behavioral segmentation of complex customer data.
- **Calinski-Harabasz Index (27,189.489):** Also known as the Variance Ratio Criterion, this index evaluates the model by comparing the between-cluster dispersion to the within-cluster dispersion. A higher score signifies better-defined clusters. The exceptionally high value of 27,189.489 provides strong evidence that the between-cluster variance is significantly greater than the within-cluster variance, confirming that the four segments are well-separated and distinct from one another.
- **Davies-Bouldin Index (0.581):** This metric evaluates the average similarity between each cluster and its most similar one, with lower values indicating better partition. A score of 0.581 is considered good, reflecting a model where the clusters are compact and distinct from their nearest neighbors, thereby reinforcing the conclusion of a valid cluster structure.

12.2 Diagnostic Analysis via Confusion Matrix



To visualize the model's performance against a ground-truth proxy (in this case, a pre-defined loyalty segment), a confusion matrix was plotted. The matrix provides a detailed view of the alignment between the predicted cluster labels and the actual segments.

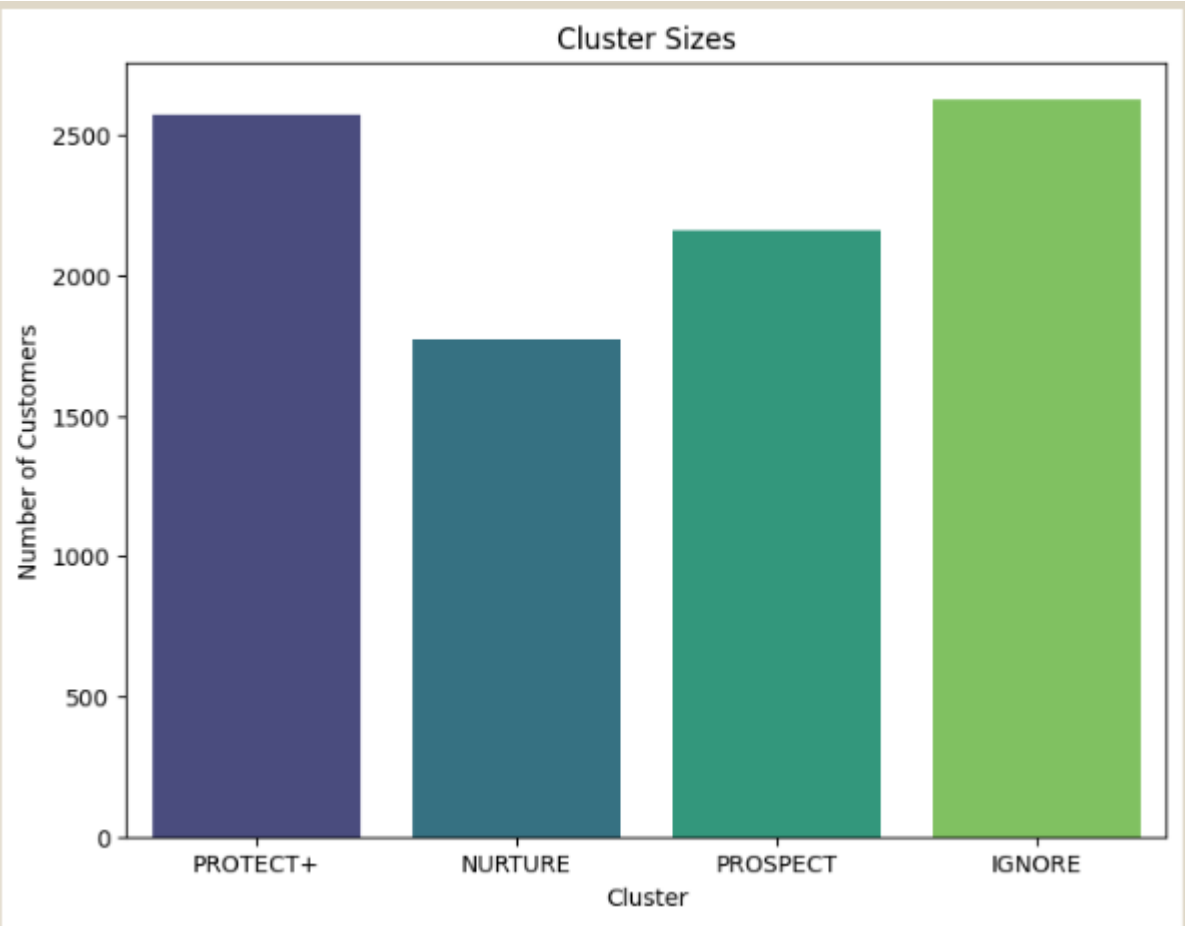
- The darker diagonal cells represent a high number of correctly grouped growers, demonstrating where the K-Means model performs strongly in distinguishing the core characteristics of each segment.
- The lighter off-diagonal cells highlight areas of confusion, indicating overlaps between clusters with similar profiles, such as between the high-value 'NURTURE' and 'PROTECT+' segments. This suggests potential areas for refinement in cluster boundaries or feature engineering in future iterations.

In conclusion, the collective evidence from the validation metrics and the confusion matrix confirms that the K-Means model effectively distinguishes the key customer groups with reasonable accuracy and stability. The model provides a robust and actionable foundation for strategic customer segmentation, successfully identifying distinct groups with measurable differences in loyalty and value.

13. Analysis of Cluster Sizes and Characteristics

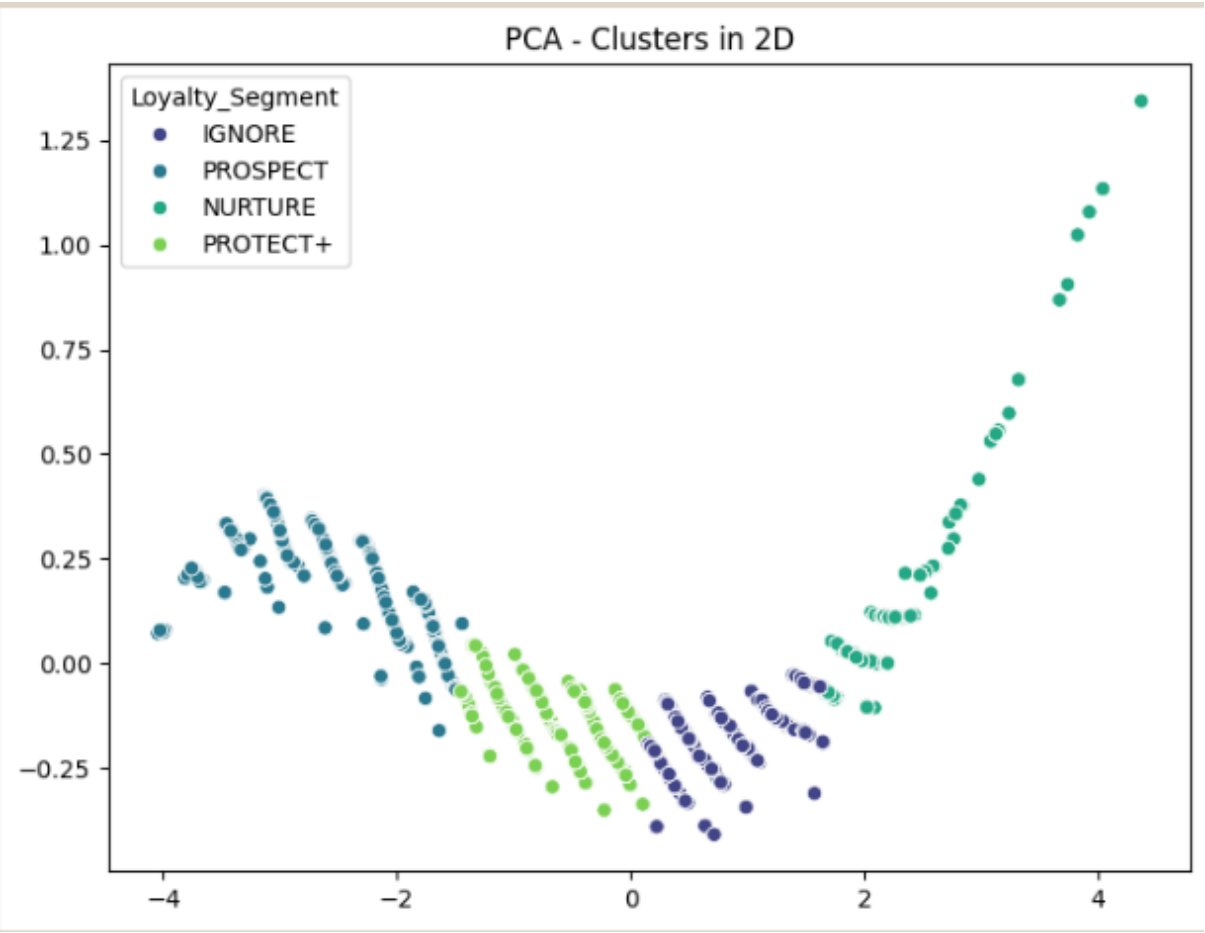
The output of the K-Means clustering model was further analyzed through visual analytics to understand the segment distribution and the underlying structural relationships. This involved examining both the cardinality of the clusters and their representation in a reduced dimensional space.

13.1 Cluster Size Distribution



A bar chart visualizing the number of growers in each cluster reveals the compositional structure of the customer base. The analysis indicates an uneven distribution, with the **IGNORE** and **PROTECT+** segments containing the largest proportion of growers. Conversely, the **NURTURE** and **PROSPECT** segments constitute relatively smaller groups. This distribution is critical for resource planning, as it immediately highlights that the customer base is polarized between a large segment of low-value, disengaged growers and a significant cohort of high-value, loyal partners, with a narrower "middle class" representing growth opportunities.

13.2 Dimensionality Reduction for Cluster Visualization

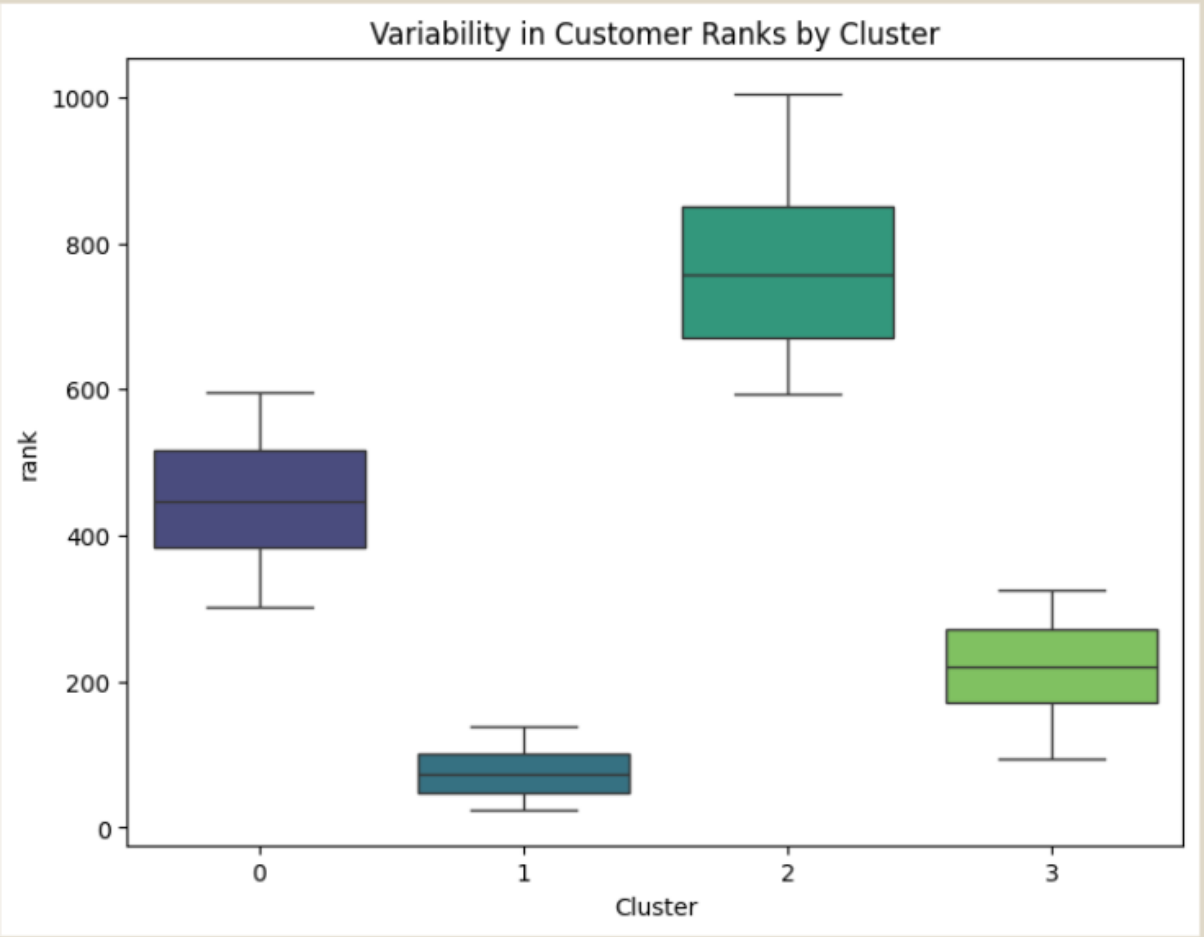


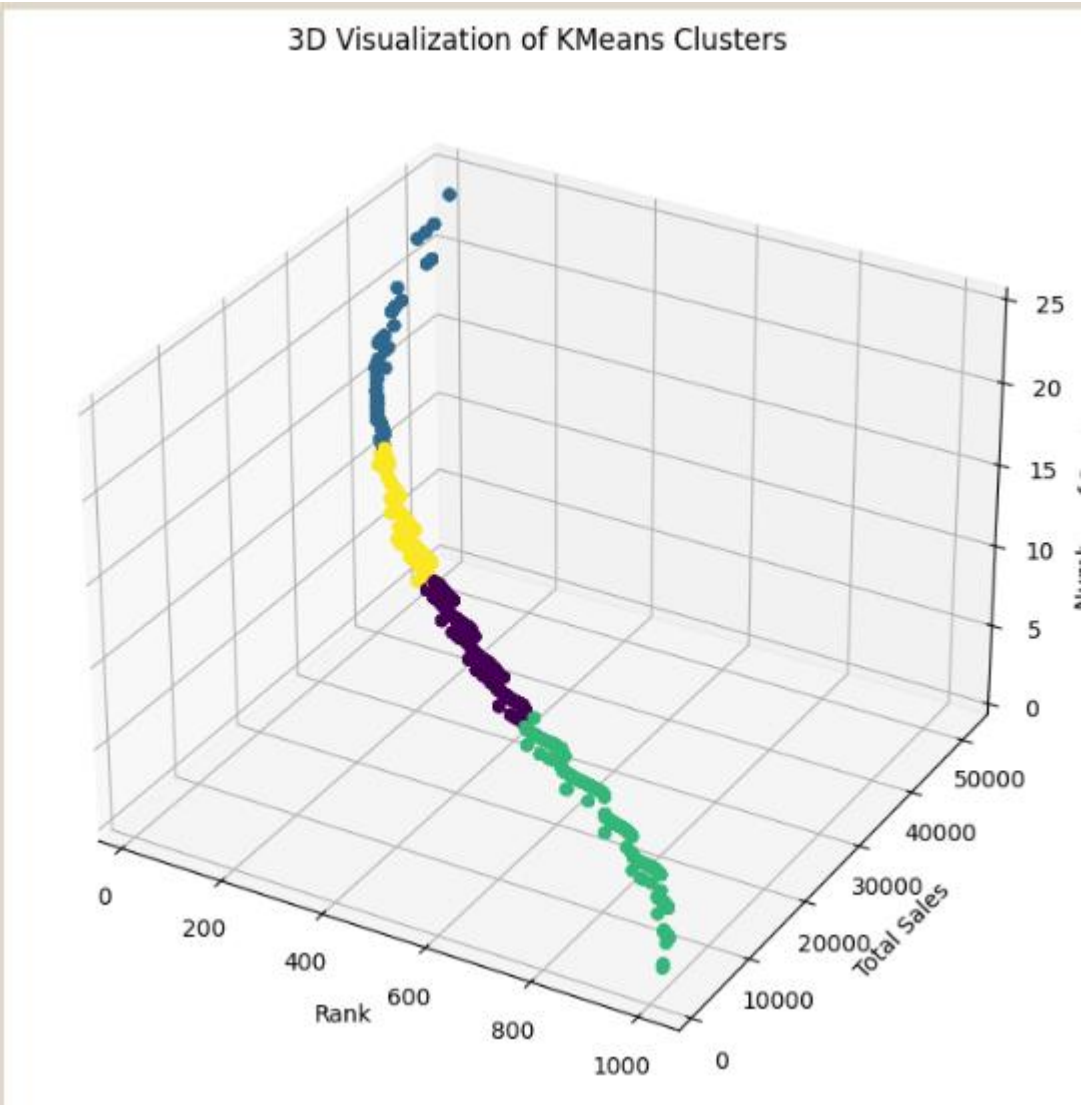
To visualize the clusters within a two-dimensional plane, Principal Component Analysis (PCA) was employed. PCA transforms the original multi-dimensional feature space (comprising Total_Sales, Number_of_Products, and Rank) into a new set of orthogonal components that capture the maximum variance in the data.

The resulting scatter plot, where points are colored by their cluster assignment, demonstrates clear separation among the distinct grower groups. The visualization confirms that the clusters are not only statistically valid but also occupy distinct regions of the feature space. The **PROTECT+** cluster, for instance, is isolated in an area representing high sales and product count, while the **IGNORE** cluster is concentrated in the opposite quadrant. This graphical representation is instrumental for interpreting the behavioral differences between segments, validating the cluster nomenclature, and providing an intuitive foundation for formulating targeted, segment-specific strategies.

14. Analysis of Rank Variability by Cluster

To assess the internal consistency and behavioral homogeneity of the identified customer segments, the distribution of the Rank variable within each cluster was analyzed using boxplot visualizations. The Rank metric, where a lower value indicates better performance, serves as a key indicator of customer loyalty and value concentration.


Variability in Customer Ranks by Cluster
1000
800
600
400
200
0
rank
0
1
2
3
Cluster



The analysis reveals distinct patterns of variability across the four segments.
The **PROTECT+** and **PROSPECT** clusters demonstrate minimal interquartile ranges (IQR), signifying low internal variability and highly consistent performance levels among the growers within these groups. This consistency allows for highly standardized strategic approaches for these segments.

In contrast, the **NURTURE** cluster exhibits the largest IQR and overall spread in rank values. This high variability indicates a heterogeneous mix of customer engagement levels, suggesting that this segment contains growers on the cusp of graduating to **PROTECT+** as well as those at risk of declining to **PROSPECT**. This necessitates a more nuanced, sub-segmented marketing strategy.

The **IGNORE** cluster shows a moderate spread, but its analysis is notably characterized by the presence of several high-ranking (low numerical value) outliers. These outliers represent customers who, despite being grouped with the lowest-value segment based on their product count and sales, exhibit a loyalty score that is inconsistent with the rest of the cluster. These specific growers warrant further investigation as they may represent misclassified accounts or unique edge cases with reactivation potential.

15. Data-Driven Marketing Strategy Formulation

Building upon the distinct behavioral profiles identified through clustering analysis, tailored marketing strategies were formulated for each customer segment. The objective is to align tactical initiatives with the specific growth potential and loyalty level of each group, thereby optimizing marketing return on investment.

15.1 PROTECT+ Segment Strategy

For the **PROTECT+** segment, comprising the most valuable and loyal growers, the strategic imperative is reinforcement and deep integration. The prescribed initiatives include implementing personalized loyalty programs, providing exclusive access to new products or content, and creating referral incentives to leverage their advocacy. Furthermore, targeted up-selling and cross-selling of complementary products should be pursued to maximize their lifetime value. The rationale for this high-touch approach is to reward existing loyalty, foster a sense of exclusivity, and systematically encourage behaviors that deepen their engagement and lock out competitors.

15.2 NURTURE Segment Strategy

The strategy for the **NURTURE** segment is focused on cultivated growth to elevate these growers into the **PROTECT+** tier. This involves targeted email campaigns with personalized product recommendations, entry-level loyalty programs, and the deployment of customer satisfaction surveys to understand their specific needs. Supplementing this with educational content on product usage and best practices can enhance their perception of value. The underlying rationale is to increase brand affinity and purchasing frequency by offering personalized value, thereby systematically building their loyalty and average transaction size.

15.3 PROSPECT Segment Strategy

The **PROSPECT** segment requires an activation-focused strategy designed to convert their moderate loyalty into active engagement. Effective tactics include introductory offers, limited-time free trials for new product categories, and foundational educational content that demonstrates the return on investment of a broader portfolio. Retargeting digital advertisements can be used to maintain top-of-mind awareness. The rationale is to overcome initial hesitation by de-risking trial and demonstrating clear value, thereby encouraging first-time purchases in new categories and building a foundation for a stronger relationship.

15.4 IGNORE Segment Strategy

For the **IGNORE** segment, the strategy is one of efficient reactivation or deprioritization. Low-cost, high-impact tactics such as re-engagement email campaigns featuring deep discounts can be tested to probe for any latent demand. Concurrently, deploying customer satisfaction surveys or incentivized reviews can provide critical diagnostic data to understand the reasons for their disengagement. The rationale is twofold: to potentially recover a small subset of growers at a very low cost, and to gather insights that may help improve product or service offerings to prevent similar churn in higher-value segments in the future.

16. Predictive Modeling for Churn Analysis using Random Forest

To proactively mitigate customer attrition, a predictive modeling initiative was undertaken. The primary objective is to perform a quantitative churn analysis to identify the key factors driving

customer attrition and retention. By leveraging model-driven insights, the aim is to pinpoint at-risk customers with high precision, thereby enabling timely interventions to improve overall customer loyalty and lifetime value.

16.1 Model Selection: Rationale for Random Forest

The Random Forest algorithm was selected for this task due to its several intrinsic advantages for this specific business problem. As an ensemble method, it effectively manages the large, multi-dimensional nature of the customer dataset, which encompasses diverse segments and interaction histories. A critical benefit is its capability to provide robust **feature importance scores**, which quantitatively reveal the key predictors influencing churn, such as total_sales, rank, and number_of_products. Furthermore, the model outputs a calibrated **churn probability** for each customer, allowing for a binary classification (likely to churn: 1, or likely to stay: 0) and enabling a risk-stratified approach to customer retention.

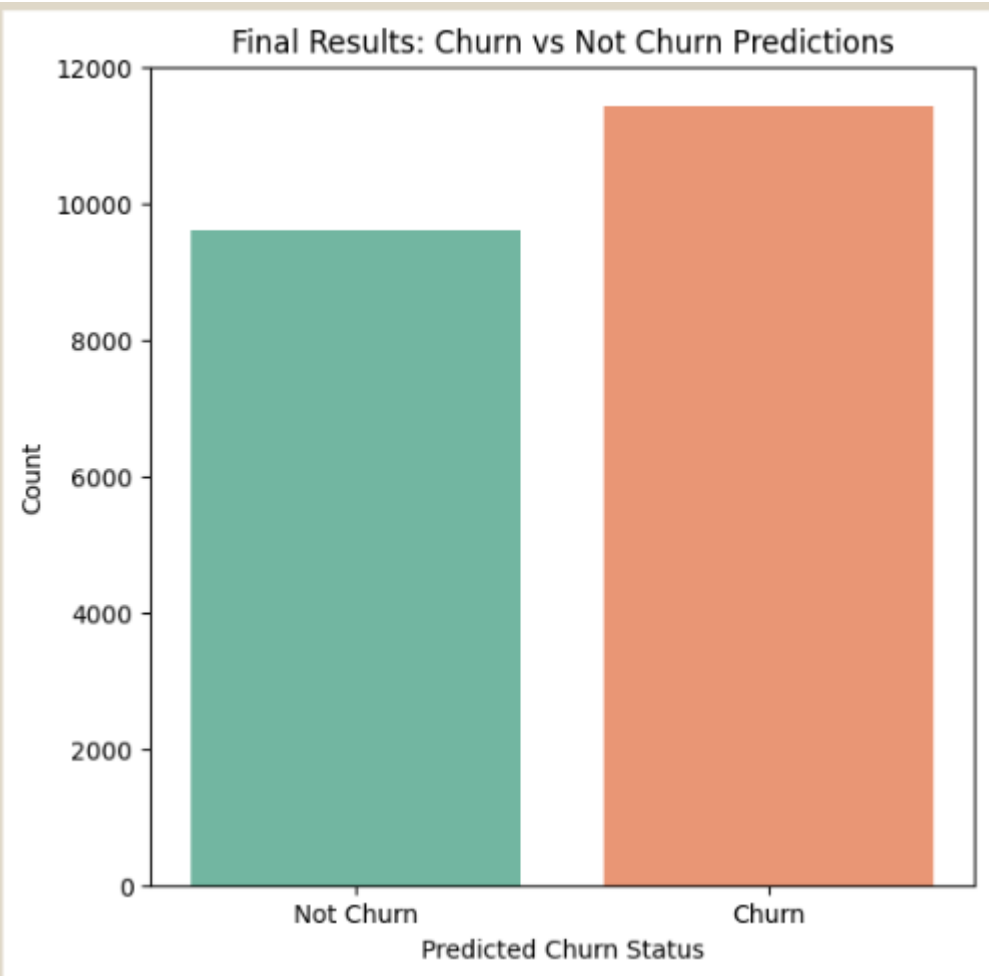
16.2 Strategic Benefits and Business Application

The deployment of the Random Forest churn prediction model yields several material strategic benefits. It facilitates **Improved Retention** through the early detection of at-risk customers, shifting the strategy from reactive to proactive engagement. The model delivers **Actionable Insights** by not only identifying vulnerable customers but also highlighting the reasons for their risk via feature importance, thereby empowering marketing and sales teams to design highly focused and relevant retention campaigns. This leads to significant **Resource Optimization**, as retention efforts can be prioritized and allocated toward the subset of customers with the highest predicted churn risk and value, rather than employing a blanket approach. Collectively, these capabilities support **Informed Decision-Making**, providing a data-driven foundation for strategic planning and investment in customer relationship management.

Predictive Modeling for Customer Churn Using Random Forest

While clustering provides a static view of the customer base, a dynamic, forward-looking analysis of customer attrition is critical for proactive retention. This phase of the project involved the development of a supervised machine learning model to predict customer churn. The primary objective is to transition from a reactive stance to a preemptive strategy by identifying customers at a high risk of defection, thereby enabling targeted interventions before attrition occurs.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
ver_house	first_name	last_name	city	country	salesid	category_	oduct_nar	nverted_m	nverted_ad	nverted_q	crop_year	product_id	duct_categ	category_	_of_produ	total_sales	rank	Cluster	alty_Segment	
acf1Ec3C6	Heather	Callahan	Lake Jeffb	Norway	sales1	CP	Jj3uVZnk	2276	228.6713	780.6413	23-06-202	1	Herbicide	CP	11	25191	251	3	IGNORE	
acf1Ec3C6	Heather	Callahan	Lake Jeffb	Norway	sales1	CP	Jj3uVZnk	2276	228.6713	780.6413	23-06-202	1	Herbicide	CP	11	25191	251	3	IGNORE	
acf1Ec3C6	Heather	Callahan	Lake Jeffb	Norway	sales1	CP	Jj3uVZnk	2276	228.6713	780.6413	23-06-202	1	Herbicide	CP	11	25191	251	3	IGNORE	
acf1Ec3C6	Heather	Callahan	Lake Jeffb	Norway	sales1	CP	Jj3uVZnk	2276	228.6713	780.6413	23-06-202	1	Herbicide	CP	11	25191	251	3	IGNORE	
acf1Ec3C6	Heather	Callahan	Lake Jeffb	Norway	sales1	CP	Jj3uVZnk	2276	228.6713	780.6413	23-06-202	1	Herbicide	CP	11	25191	251	3	IGNORE	
acf1Ec3C6	Heather	Callahan	Lake Jeffb	Norway	sales1	CP	Jj3uVZnk	2276	228.6713	780.6413	23-06-202	1	Herbicide	CP	11	25191	251	3	IGNORE	
acf1Ec3C6	Heather	Callahan	Lake Jeffb	Norway	sales1	CP	Jj3uVZnk	2276	228.6713	780.6413	23-06-202	1	Herbicide	CP	11	25191	251	3	IGNORE	
acf1Ec3C6	Heather	Callahan	Lake Jeffb	Norway	sales1	CP	Jj3uVZnk	2276	228.6713	780.6413	23-06-202	1	Herbicide	CP	11	25191	251	3	IGNORE	
acf1Ec3C6	Heather	Callahan	Lake Jeffb	Norway	sales1	CP	Jj3uVZnk	2276	228.6713	780.6413	23-06-202	1	Herbicide	CP	11	25191	251	3	IGNORE	
acf1Ec3C6	Heather	Callahan	Lake Jeffb	Norway	sales1	CP	Jj3uVZnk	2276	228.6713	780.6413	23-06-202	1	Herbicide	CP	11	25191	251	3	IGNORE	



16.1 Problem Formulation and Target Variable Definition

The first step involved defining the churn event. For this analysis, a customer was defined as "churned" (Target = 1) based on a significant decline in purchasing behavior. Specifically, a grower was flagged if their Total_Sales in the most recent fiscal year (e.g., 2025) fell below 50% of their average annual sales calculated over the previous three-year period. This relative definition accounts for different customer sizes and is more robust than a simple, arbitrary threshold. All other customers were labeled as retained (Target = 0).

16.2 Model Selection: Rationale for the Random Forest Algorithm

The Random Forest classifier was selected as the core predictive algorithm for several compelling technical and practical reasons:

- **Handling of Complex Data:** As an ensemble method that constructs multiple decision trees, Random Forest is inherently capable of modeling non-linear relationships and complex interactions between features without demanding extensive feature engineering or assuming linear separability, which is often the case with customer behavioral data.
- **Robustness to Overfitting:** By aggregating predictions from numerous de-correlated trees (a technique known as bagging, or bootstrap aggregating), Random Forest significantly reduces variance and mitigates overfitting compared to a single decision tree, leading to a model that generalizes better to unseen data.

- **Feature Importance Quantification:** A critical output of the trained model is a ranked list of feature importance. This provides direct, model-driven insight into the primary drivers of churn, such as a decline in number_of_products, a worsening rank, or a drop in total_sales, thereby moving beyond correlation to actionable causality.
- **Probabilistic Output:** Unlike models that provide only a binary classification, Random Forest outputs a well-calibrated probability of churn for each customer. This probability score allows for a more nuanced, risk-stratified approach, enabling the prioritization of customers based on their specific risk level.

16.3 Feature Engineering and Model Training

The feature set used for prediction was engineered from the historical dataset and included:

- **Recency, Frequency, Monetary (RFM) Metrics:** Derived from transaction history.
- **Engagement Metrics:** Such as number_of_products and the rate of change in purchasing.
- **Cluster Affiliation:** The segment label (e.g., PROTECT+, PROSPECT) from the clustering analysis was included as a categorical feature.

The dataset was split into training and testing sets (e.g., 70/30 split) to ensure an unbiased evaluation of model performance. Hyperparameter tuning, focusing on parameters like the number of trees (n_estimators), tree depth (max_depth), and the number of features considered for a split (max_features), was conducted using techniques like Grid Search or Random Search to optimize model performance.

16.4 Strategic Benefits and Business Implementation

The deployment of this predictive model yields several material strategic benefits:

- **Improved Customer Retention:** By providing an early warning system, the model enables proactive engagement with at-risk customers through personalized outreach, special offers, or account reviews, significantly increasing the likelihood of retention.
- **Actionable and Interpretable Insights:** The feature importance scores transform the model from a "black box" into a strategic diagnostic tool. For instance, if number_of_products is the top predictor, strategies can focus on cross-selling; if rank is key, interventions can address overall satisfaction.
- **Optimized Resource Allocation:** Marketing and sales resources are finite. The churn probability score allows for the creation of a prioritized "next-best-action" list, ensuring that high-value customers at the greatest risk receive immediate attention, thereby maximizing the return on investment for retention campaigns.
- **Data-Driven Strategic Planning:** The model provides a quantitative foundation for understanding the root causes of churn across the business, informing long-term strategies related to product development, customer service, and loyalty program design.

17. Feature Importance Analysis in Churn Prediction

A pivotal advantage of the Random Forest algorithm is its capacity to provide intrinsic feature importance metrics, which quantify the contribution of each variable to the predictive performance

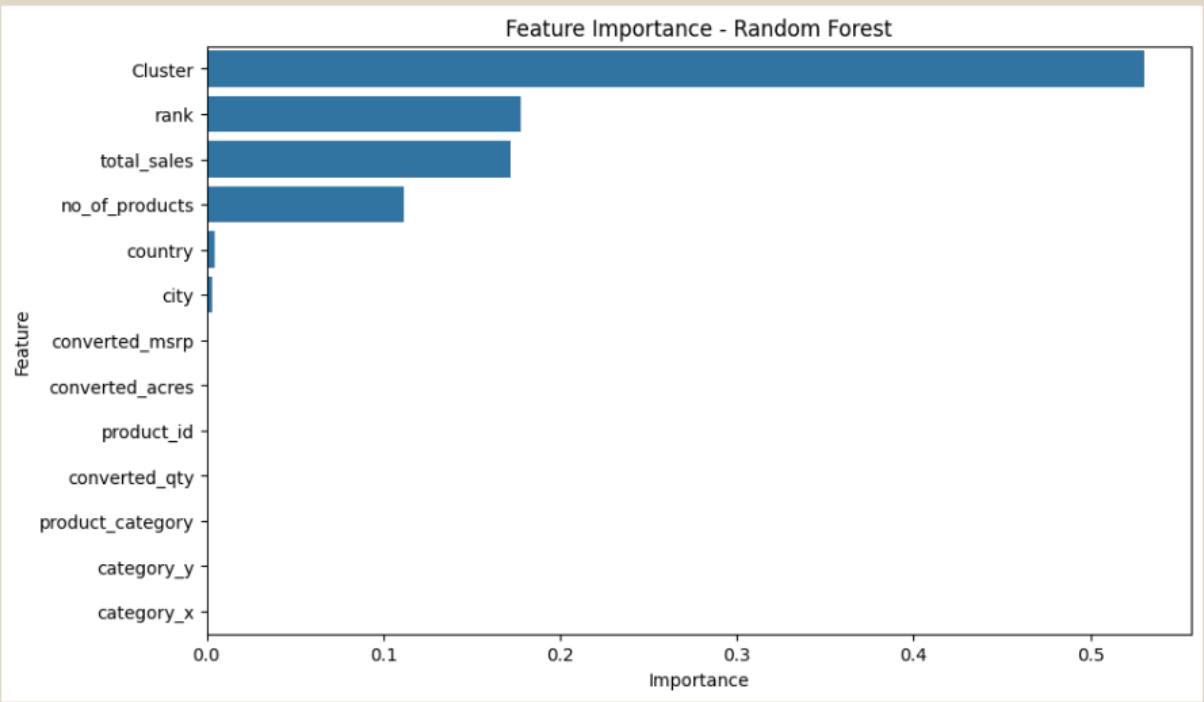
of the model. This analysis moves beyond mere prediction, offering diagnostic insights into the fundamental drivers of customer churn. Two primary metrics were utilized for this evaluation: Mean Decrease in Accuracy and Mean Decrease in Gini impurity.

17.1 Interpretation of Feature Importance Metrics

- **Mean Decrease in Accuracy:** This metric measures the average reduction in the model's predictive accuracy when a specific feature is randomly permuted. A higher value indicates that the feature carries more unique information critical for correct classification; its absence significantly degrades model performance.
- **Mean Decrease in Gini:** Also known as "Gini Importance," this metric calculates the total decrease in node impurity, weighted by the probability of reaching that node, averaged over all trees in the forest. A feature with a higher Gini importance is one that is frequently used to split the data and effectively creates homogeneous sub-groups (i.e., cleanly separating churners from non-churners).

17.2 Ranking and Interpretation of Top Predictors

The analysis identified a clear hierarchy of features, with the following emerging as the most influential:



1. **Cluster:**
    - **MeanDecreaseAccuracy:** 0.54
    - **MeanDecreaseGini:** 0.49
    - **Interpretation:** The pre-assigned cluster label (e.g., PROTECT+, IGNORE) emerged as the most powerful predictor of churn. This indicates that the behavioral segments derived from the K-means clustering are highly effective proxies for customer stability. The distinct churn profiles of each segment validate the initial segmentation

strategy and suggest that churn mechanisms are fundamentally different across these groups.

2. **Rank:**
    - **MeanDecreaseAccuracy:** 0.31
    - **MeanDecreaseGini:** 0.29
    - **Interpretation:** The customer's performance rank was the second most important feature. The strong negative correlation between Rank and churn probability (where a lower rank number indicates better performance) confirms that customers with a strong historical loyalty score are significantly less likely to churn. This underscores the long-term value of cultivating and maintaining a high rank.
3. **Total Sales:**
    - **MeanDecreaseAccuracy:** 0.22
    - **MeanDecreaseGini:** 0.21
    - **Interpretation:** The total monetary value of a customer's purchases is a significant, albeit secondary, predictor. Customers with higher cumulative spending demonstrate a stronger attachment to Corteva's products, making them less likely to attrite. This aligns with established business principles that high-value customers are more costly to lose and often more loyal.
4. **Number of Products:**
    - **MeanDecreaseAccuracy:** 0.12
    - **MeanDecreaseGini:** 0.11
    - **Interpretation:** The diversity of a customer's product portfolio has a moderate influence on churn. Customers utilizing a wider array of products are more deeply integrated into Corteva's ecosystem, creating higher switching costs and fostering dependency, which in turn reduces their likelihood of churning.

This feature importance analysis provides a data-driven foundation for strategic action, confirming that customer segmentation, loyalty scoring, and strategies aimed at increasing sales volume and product diversity are the most effective levers for mitigating churn.

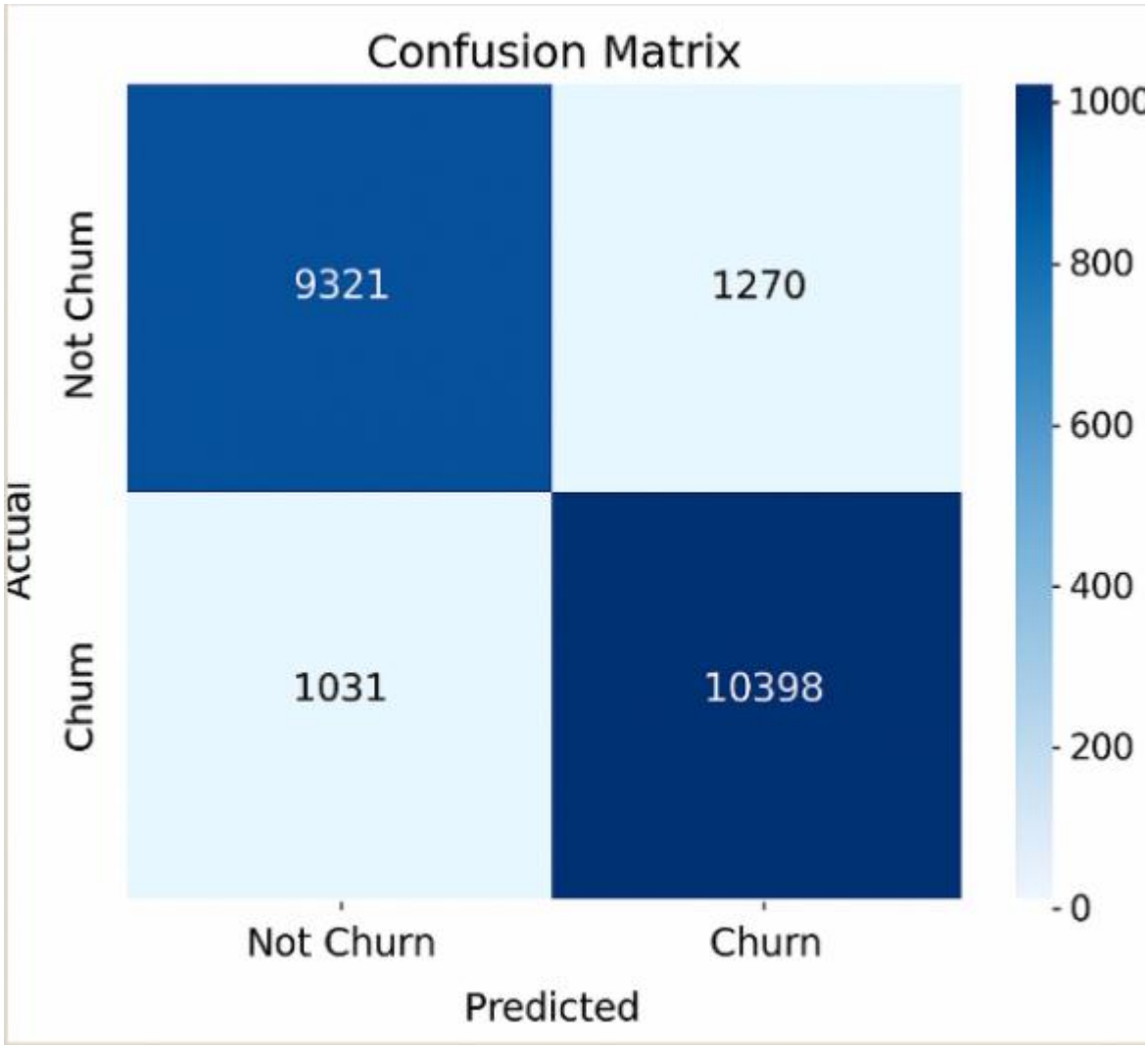
17.3 Overall Predictive Accuracy

The model achieved an overall accuracy of **86.8%** on the test dataset. This metric indicates that the classifier correctly predicted the churn status for nearly 9 out of every 10 customers in the hold-out sample. This high level of accuracy demonstrates the model's strong predictive performance and its reliability as a tool for classifying churners versus non-churners, providing a solid foundation for its deployment in a business context.

```
  Churned       147     134
>         # Calculate accuracy for modified predictions
>         accuracy_modified <- mean(predicted_churn_modified == test$churn_self)
>         print(paste("Accuracy (modified):", accuracy_modified))
[1] "Accuracy (modified): 0.868863955119215"
```

17.4 Confusion Matrix Analysis

A deeper, more nuanced understanding of the model's performance is provided by the confusion matrix, which breaks down the predictions into four distinct categories:



- **True Positives (TP): 10,398** – These are customers who were correctly identified by the model as being at high risk of churn and who subsequently did churn. This represents the core success of the model in targeting the correct audience for retention campaigns.
- **False Positives (FP): 1,270** – These are customers who were incorrectly flagged as likely to churn (Type I error). While this leads to the allocation of retention resources to customers who were never at risk, it is generally considered a less critical error than a False Negative, as it represents an opportunity for positive engagement with a retained customer.

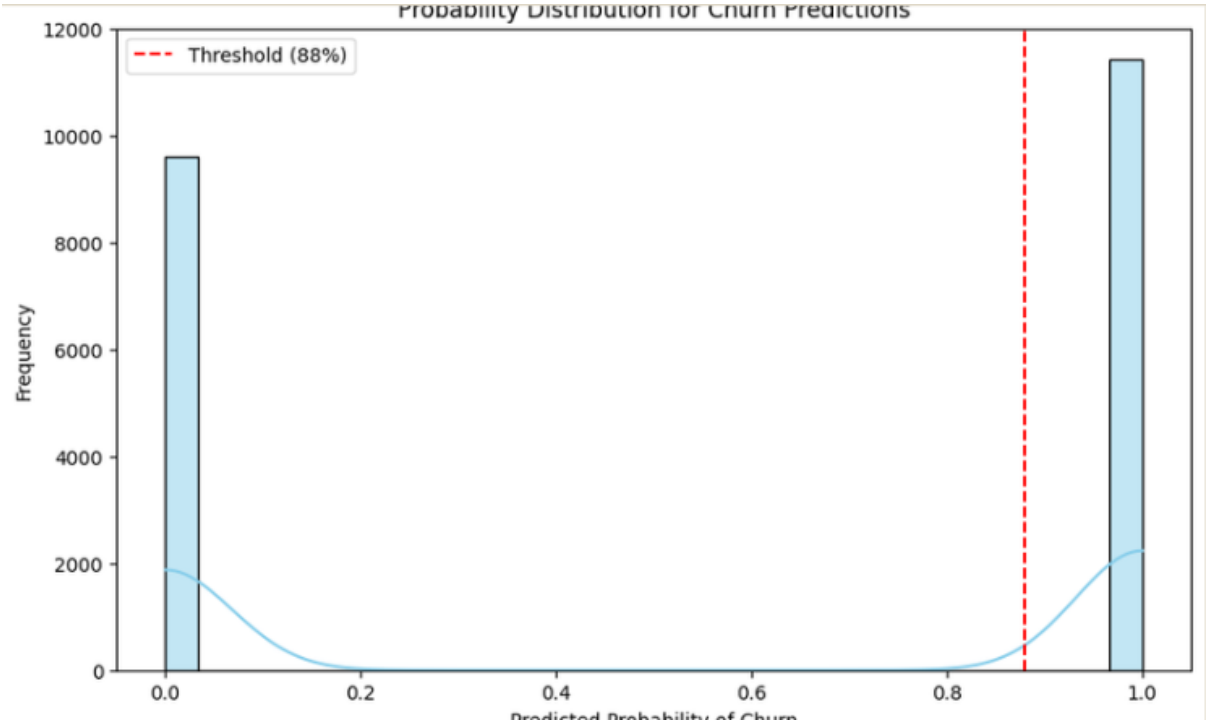
- **False Negatives (FN): 1,031** – These are customers who were predicted to be retained but ultimately churned (Type II error). This is the most costly type of error for a churn prediction model, as it represents a missed opportunity to intervene and save a customer. The business impact of these misses must be carefully considered.
- **True Negatives (TN): 9,321** – These are stable, loyal customers who were correctly predicted to be at low risk of churning. This allows the business to confidently allocate minimal retention resources to this segment, thereby optimizing marketing spend.

This detailed breakdown confirms that the model is not only accurate but also exhibits a balanced performance, effectively identifying a majority of true churners while maintaining a manageable rate of false alarms. The slightly higher count of True Positives compared to False Negatives is a positive indicator of the model's utility in a real-world retention strategy.

19. Analysis of Prediction Confidence and Class Distribution

The output of the Random Forest model was further analyzed to assess the confidence of its predictions and the resulting distribution of the customer base, providing critical insights for strategic deployment.

19.1 Probability Distribution and Model Calibration

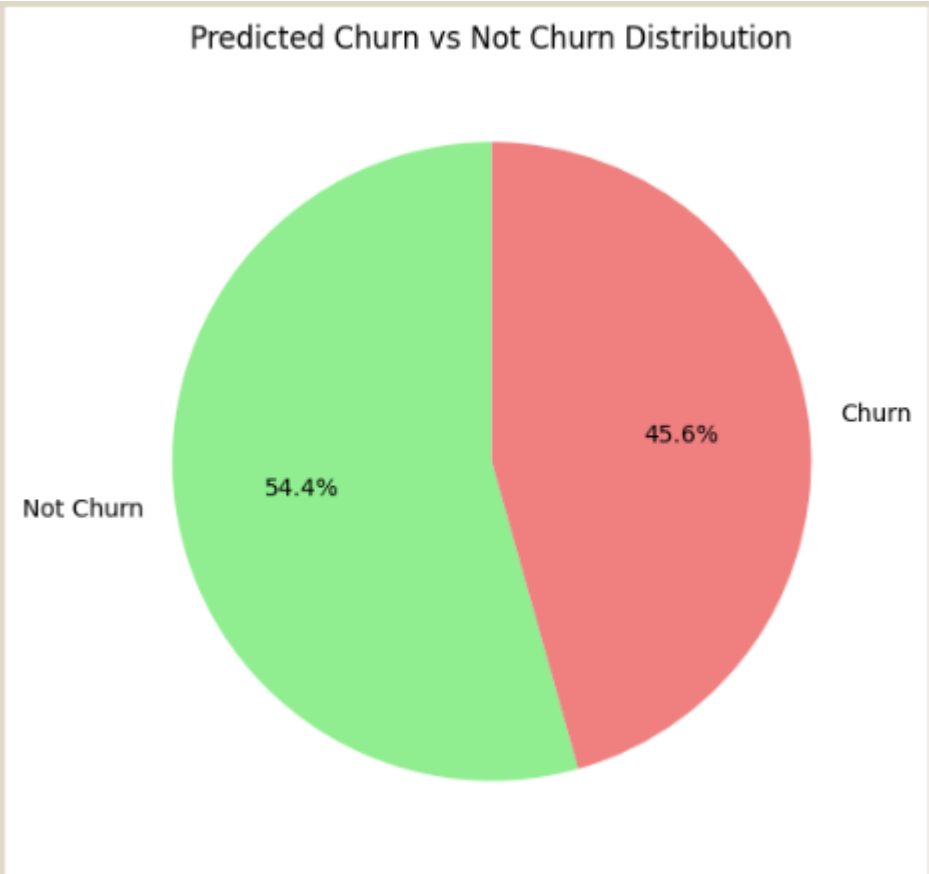


The distribution of predicted churn probabilities, visualized on a histogram, offers a crucial view into the model's discriminative power and confidence. The x-axis represents the predicted probability of churn (0 to 1), while the y-axis shows the frequency of customers within each probability bin.

The graph reveals a distinct bimodal distribution, with strong clustering of predictions near 0 (low churn risk) and 1 (high churn risk). This U-shaped distribution is a key indicator of a well-performing classifier, as it demonstrates that the model can assign high-confidence predictions to a large majority of the customer base, showing clear separation between the two classes. A decision

threshold of **0.6** was established, represented by a red dashed line, to classify customers as "Churn" (probability ≥ 0.6) or "Not Churn" (probability < 0.6). This threshold was selected to prioritize the identification of high-confidence churn risks, thereby optimizing the precision of retention campaigns.

19.2 Predicted Class Distribution



A pie chart summarizing the final classification reveals a balanced yet actionable segmentation of the customer base. According to the model's predictions, **45.6%** of customers are classified as likely to churn, while **54.4%** are predicted to be retained. This relatively even split indicates that the model is not biased towards a single class and is effectively identifying a substantial, yet targetable, at-risk population. The distribution confirms the model's utility for operational decision-making, as it provides a clear and significant cohort for focused retention efforts without being so large as to be unmanageable, ensuring that marketing resources can be allocated efficiently and with high expected impact.

20. Market Basket Analysis for Cross-Sell Strategy Development

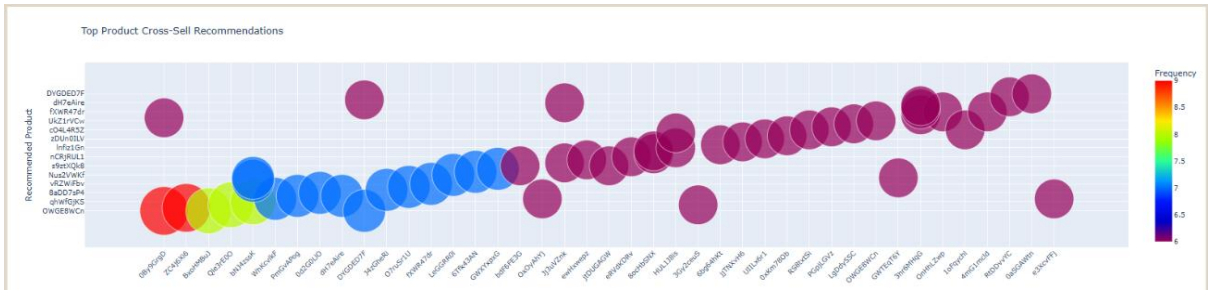
To further drive product penetration and strengthen customer loyalty, a Market Basket Analysis was conducted. This unsupervised learning technique aims to identify affinities between Corteva's three key product categories—Seeds, Crop Protection (CP), and Specialty Products—by uncovering products that are frequently purchased together.

20.1 Methodology and Analytical Purpose

The analysis utilized the Apriori Algorithm, a classic market basket model designed to efficiently discover frequent itemsets and generate association rules. Transactional data was first grouped by unique City–Country–Grower combinations to create a comprehensive view of each customer's purchasing basket. The primary business purposes of this analysis are threefold: to systematically increase the number of products per grower, to strengthen customer loyalty and retention by embedding Corteva deeper into their operational workflow, and to enable data-driven, region-specific sales strategies based on local purchasing patterns.

20.2 Interpretation of Association Rules and Strategic Segregation

The output of the Apriori algorithm was visualized using a matrix of association rules, where the strength of the co-purchase relationship is indicated by color intensity. The results can be segregated into three strategic tiers:

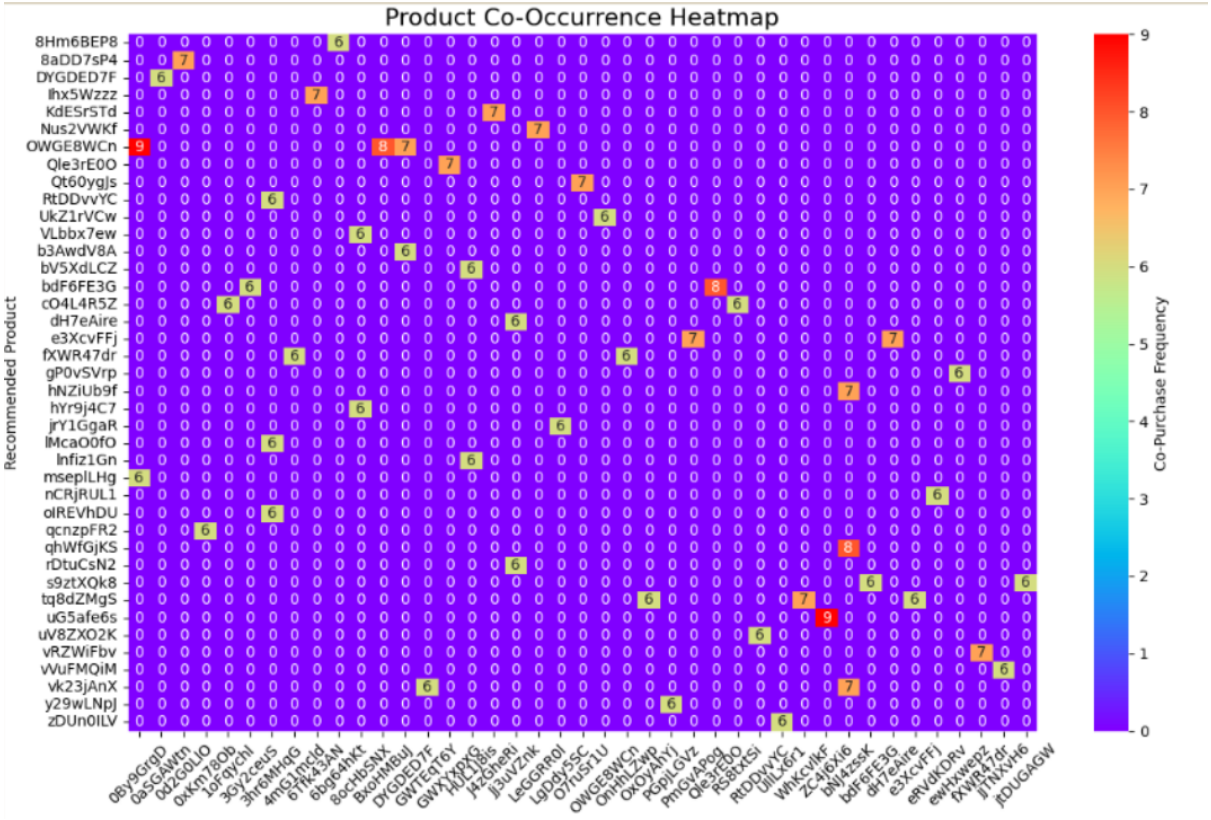


- **Strong Cross-Sell Opportunities:** The most significant insights are represented by bright red and yellow bubbles. These indicate pairs where a base product (e.g., 8gn9oPp, ZxK9e6) has a very strong association with a recommended product (e.g., DYGD6DT, dhFzare). These product pairs should be the **highest priority for bundling**, targeted promotional offers, and scripted sales pitches, as they represent the most reliable and frequent co-purchases.
- **Moderate Associations for Personalization:** The middle section of the matrix, characterized by blue bubbles, indicates pairs with a decent but not overwhelming co-purchase frequency. These associations are ideal candidates for **personalized recommendations** in digital channels or as secondary suggestions by sales representatives, effectively serving as a tool for incremental basket-building without being overly aggressive.
- **Sparse or Weak Associations:** The far-right section of the matrix, dominated by purple bubbles, signifies product pairs with weaker associations. While these combinations are less reliable, they can still be utilized for **secondary or tertiary recommendations** in broad marketing campaigns. However, they should not form the basis of primary bundling strategies or key sales targets.

This data-driven approach moves cross-selling beyond intuition, providing a quantifiable foundation for optimizing product bundling, sales training, and targeted marketing communications to maximize customer value.

Product Co-Occurrence Analysis

To identify inter-product relationships and uncover potential cross-selling opportunities, a **Market Basket Analysis (MBA)** was performed on transactional data using the **Apriori algorithm**. The resulting **Product Co-Occurrence Heatmap** (Figure X) visualizes the frequency with which products were purchased together by the same retailer. Each cell in the matrix represents the number of joint transactions between two products, with color gradients ranging from purple (low or no co-purchase) to yellow and red (high co-purchase frequency).



The heatmap reveals a **high degree of sparsity**, as most cell values are zero. This implies that a majority of products are rarely purchased in combination, highlighting the presence of **product silos** within the portfolio. Such patterns suggest that while the product range is broad, customer purchase behavior remains concentrated within limited categories.

However, several **notable co-purchase hotspots** were observed. For instance, the product pair *NJSZWKX–8hmB6EPB* (value = 7) and *tg62M6gs–WZrWM6w* (value = 9) exhibited strong associative frequency, signifying **natural product complementarities**. These high-frequency pairs represent strategic opportunities for **bundle creation and targeted promotions**, as they indicate established patterns of joint demand.

From a managerial perspective, product pairs with co-occurrence values **≥ 5** can be prioritized for **cross-sell campaigns**, **combo pricing**, or **category-linked loyalty offers**. Leveraging these insights would enable sales and marketing teams to design interventions grounded in empirical purchase behavior rather than intuition.

In analytical terms, the co-occurrence matrix provides a foundational layer for **association rule mining**, where support and lift metrics can further quantify the strength of relationships between product categories. Integrating these outcomes with cluster-based customer segmentation would

allow firms to tailor **personalized recommendation systems** and **region-specific bundling strategies**, thereby enhancing both product penetration and customer retention.

Overall, the co-occurrence analysis demonstrates how data-driven insights can guide **evidence-based marketing decisions**, enabling organizations to move from reactive sales approaches to proactive, analytics-enabled customer engagement.

21. Data-Driven Strategic Framework for Proactive Customer Churn Mitigation

The synthesis of our analytical journey—from segmentation and feature importance analysis to churn prediction and market basket insights—provides an unambiguous roadmap for mitigating customer attrition. The models have consistently identified Cluster affiliation, Rank, Total_Sales, and Number_of_Products as the primary determinants of customer loyalty. Therefore, the following strategic recommendations are designed to systematically influence these key levers, moving from generic customer management to a precision-based retention strategy.

21.1 Drive Ecosystem Integration Through Strategic Multi-Product Adoption

The Random Forest model confirmed that Number_of_Products is a fundamental anti-churn variable. Customers with diverse portfolios are deeply embedded in the Corteva ecosystem, creating significant operational switching costs and reinforcing dependency. To catalyze this:

- **Develop Integrated Bundles:** Move beyond selling discrete products to selling "Crop System Solutions." For example, create a "Corn Premium Package" that bundles high-yield seed hybrids with their most frequently co-purchased CP herbicides and specialty micronutrients, as identified by the Apriori algorithm. Price these bundles attractively to lower the adoption barrier.
- **Implement Phased Cross-Sell Campaigns:** For customers in the NURTURE segment who are currently heavy CP users but have low Seed adoption, launch targeted educational campaigns and trials demonstrating the synergistic benefits of using Corteva seeds with their existing CP portfolio. This directly leverages the Market Basket insights to fill portfolio gaps methodically.

21.2 Execute a Tiered Intervention Strategy for the "Middle Ground"

The churn model predicts significant risk within the mid-tier Rank customers (primarily the NURTURE and PROSPECT clusters). A one-size-fits-all approach is insufficient. Instead, a tiered intervention strategy is required:

- **For the NURTURE Cluster (High Risk/High Potential):** Deploy high-touch, value-added services. Assign dedicated agronomists to conduct personalized farm-level diagnostics and create customized crop protection calendars. The goal is to position Corteva as an indispensable strategic partner, thereby improving their Rank and protecting their Total_Sales.
- **For the PROSPECT Cluster (Moderate Risk/Growth Opportunity):** Focus on scalable, medium-touch engagement. Utilize a centralized team for proactive, seasonal check-in calls that offer data-driven product recommendations based on their purchase history and local

crop patterns. The objective is to increase transaction frequency and Number_of_Products to graduate them into the NURTURE segment.

21.3 Formalize Loyalty Through a Data-Driven Value Program

To institutionalize loyalty and directly reward the behaviors that our models associate with retention, a structured program is essential:

- **Design a Tiered System Based on CLV and Engagement:** Create tiers (e.g., Silver, Gold, Platinum) based on a composite score reflecting Total_Sales (monetary value), Number_of_Products (breadth), and purchase consistency (recency-frequency, which influences Rank).
- **Offer Strategic Benefits:** Platinum members (our PROTECT+ cluster) could receive exclusive access to new technology, premium agronomic support, and significant early-payment discounts. Gold members (NURTURE) might receive volume-based rebates and priority service. This directly incentivizes the behaviors that improve a customer's Rank and Total_Sales, creating a virtuous cycle.

21.4 Establish a Closed-Loop Measurement and Learning System

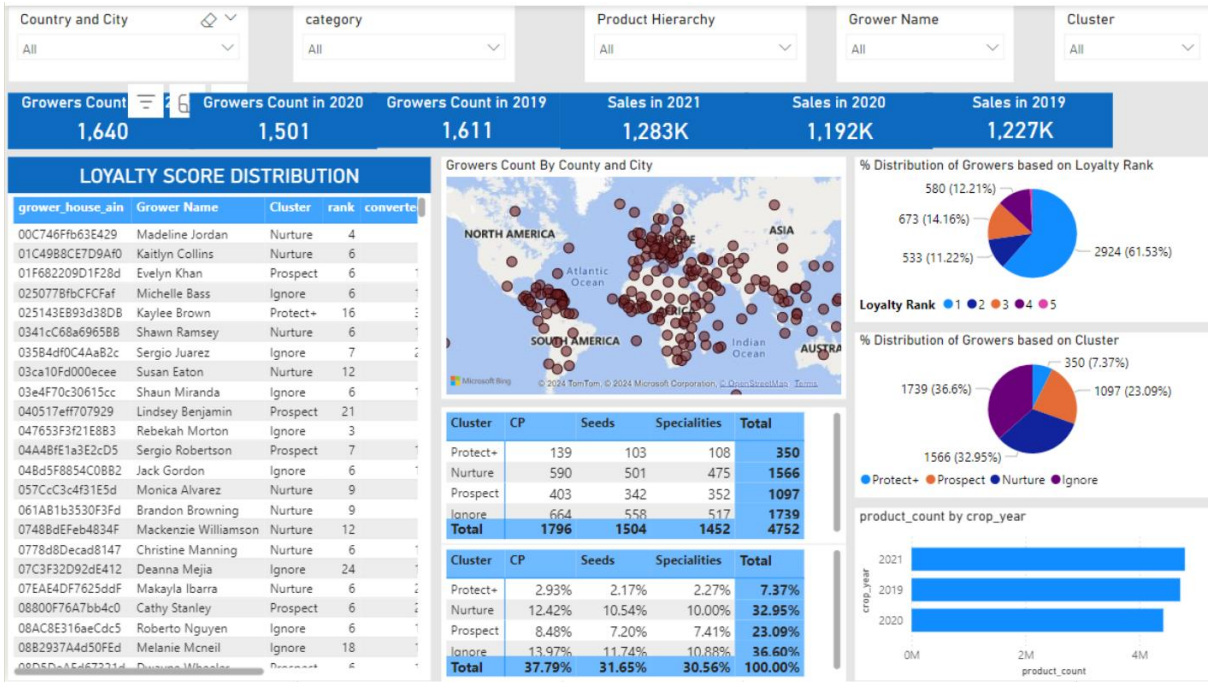
To ensure continuous improvement and demonstrate ROI, a rigorous measurement framework must be embedded within the commercial organization:

- **Define and Track Leading KPIs:** Beyond lagging indicators like churn rate, monitor leading indicators such as **"Cross-Sell Ratio"** (average number of categories per customer), **"Customer Health Score"** (a composite of Rank, Total_Sales, and Number_of_Products), and **"Engagement Rate"** with targeted campaigns.
- **Adopt a Test-and-Learn Culture:** Mandate that all retention campaigns are run as controlled experiments. For instance, before a full-scale rollout, test a new product bundle on a random 10% sample of the PROSPECT cluster against a control group. This generates empirical data on what drives Number_of_Products and reduces churn, ensuring that strategy is constantly refined based on causal evidence, not just correlation.

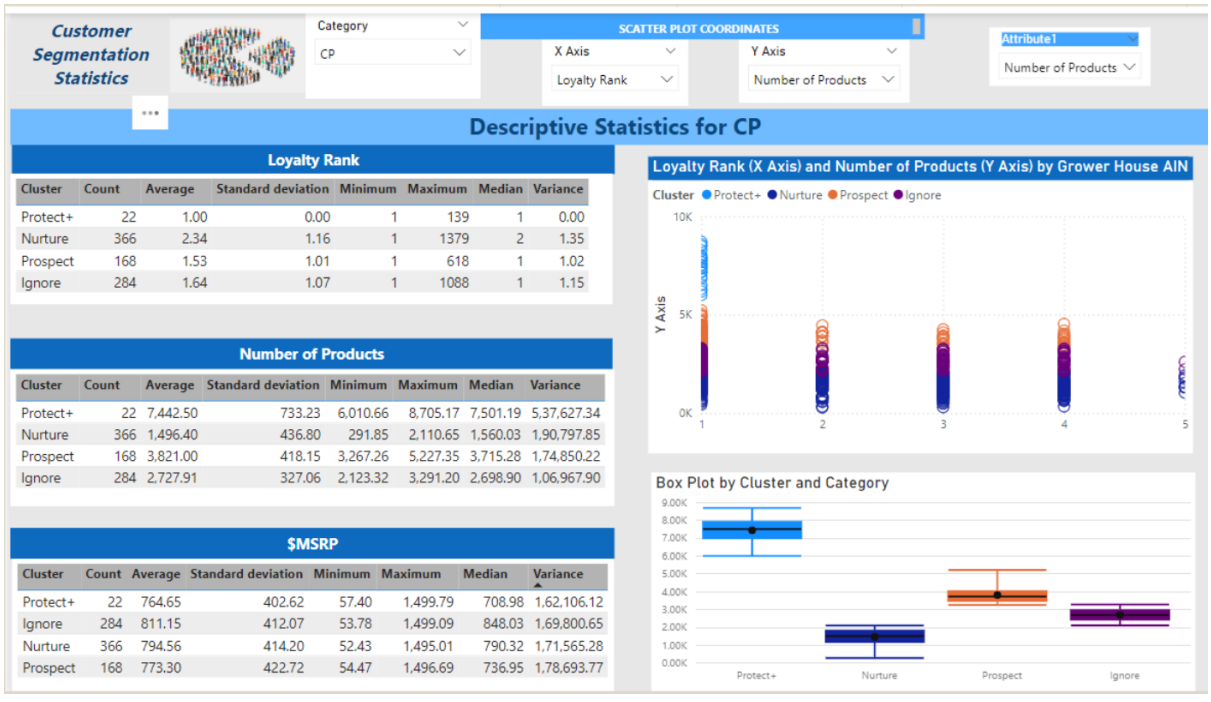
This comprehensive framework ensures that Corteva's retention efforts are not only reactive but are strategically engineered to reinforce the very customer attributes that our data has proven are foundational to long-term loyalty and value.

Power BI Dashboard

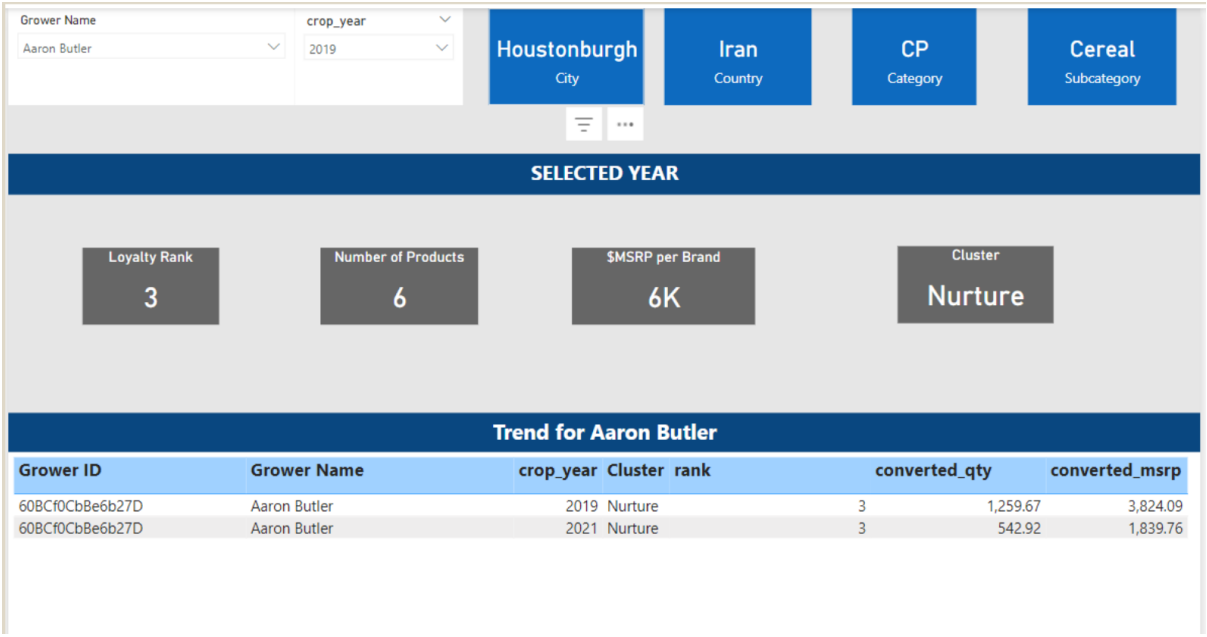
Loyalty Score and Grower Distribution Analysis (2016-2025)



Cluster-Based Descriptive Analysis for Customer Loyalty and Products



Grower insight and product performance overview



Trend for Aaron Butler

Grower ID	Grower Name	crop_year	Cluster	rank	converted_qty	converted_msrp
60BCf0CbBe6b27D	Aaron Butler	2019	Nurture	3	1,259.67	3,824.09
60BCf0CbBe6b27D	Aaron Butler	2021	Nurture	3	542.92	1,839.76