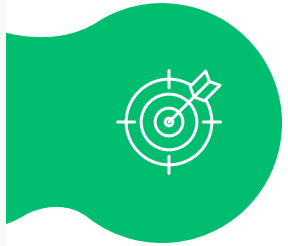
INSIGHTS FROM

INSTACART MARKET BASKET ANALYSIS

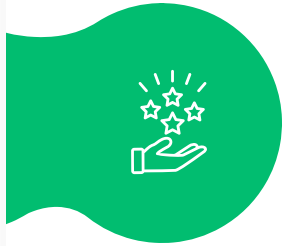
Charles Bryant

CONTEXT



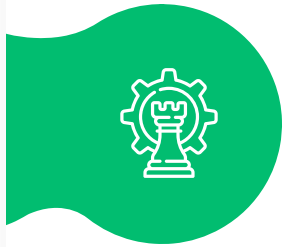
Objective

Unlock revenue potential by boosting customer engagement and retention through personalized offerings.



Opportunity

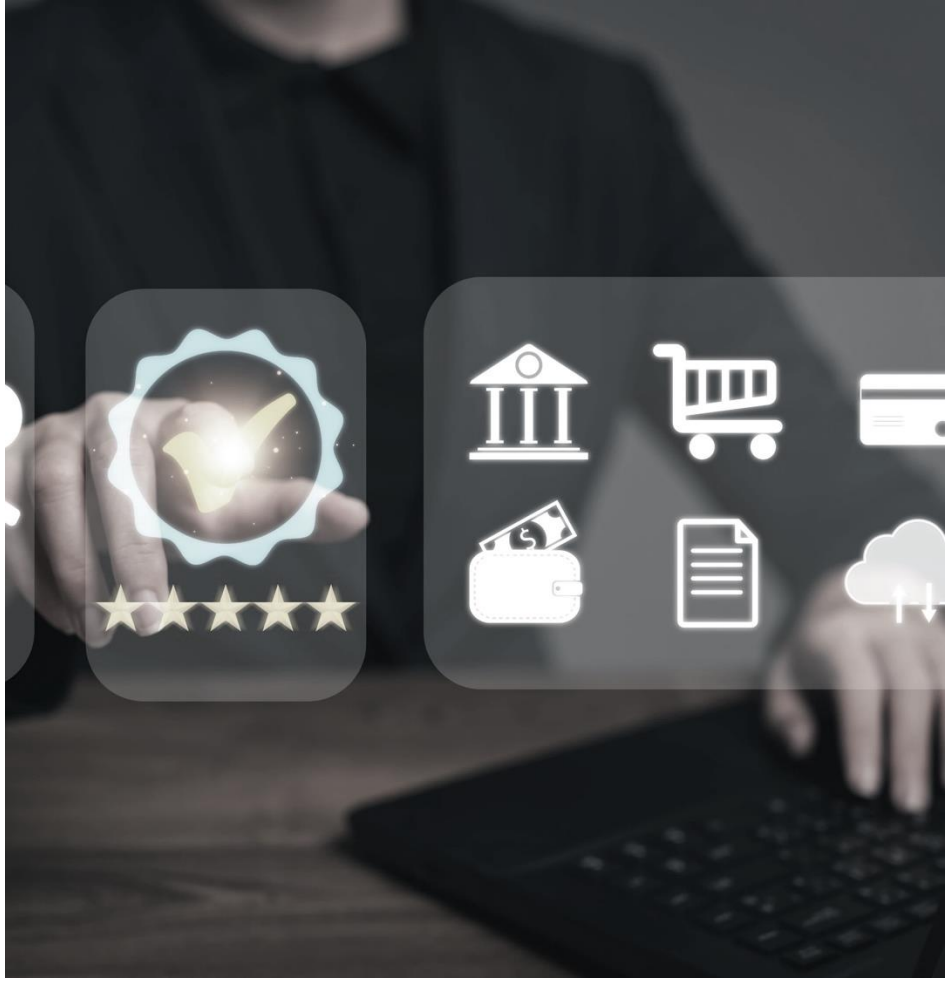
Capitalize on robust historical data to drive targeted marketing, optimize inventory, and improve operational efficiency.



Strategy

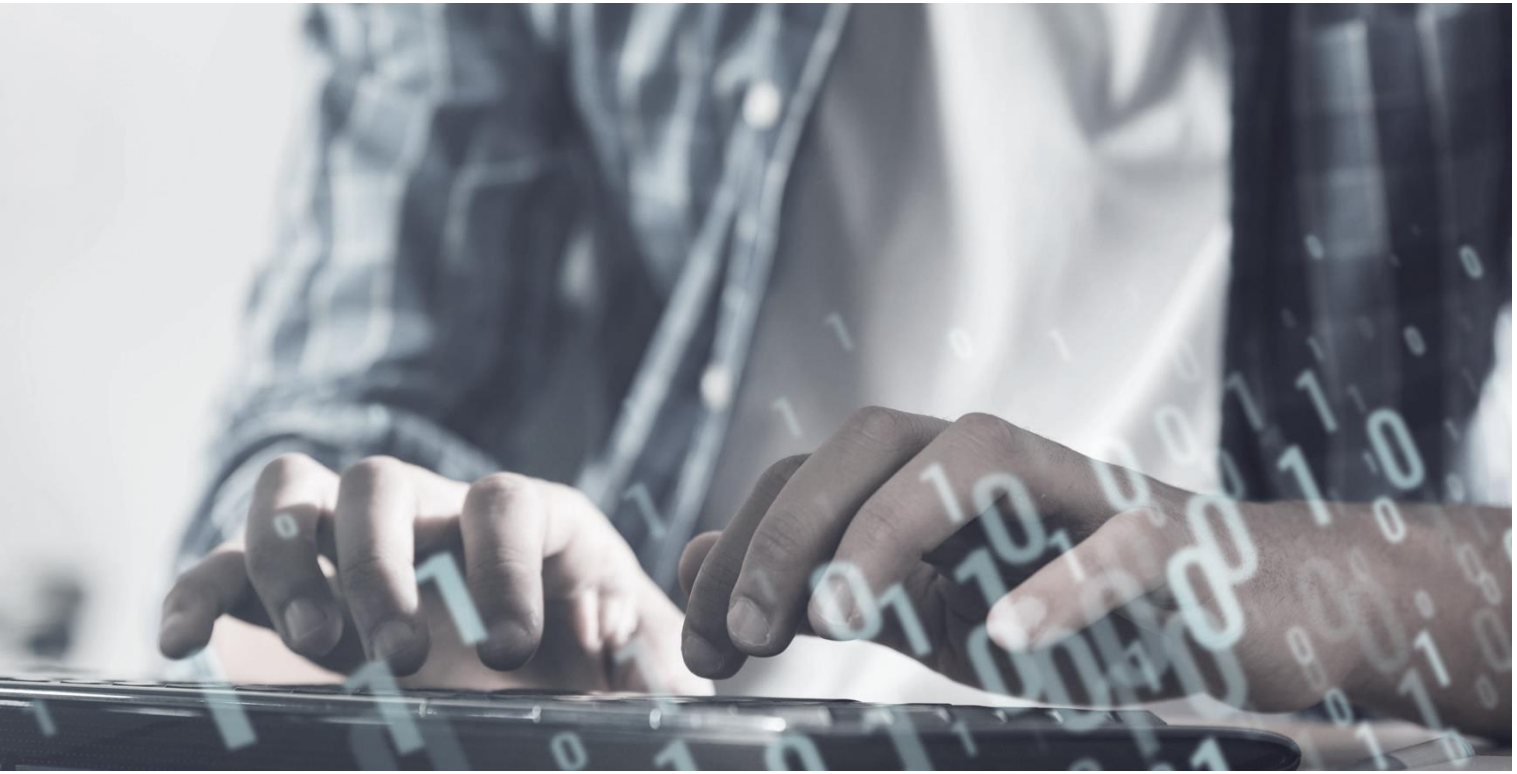
Use enterprise analytics to accelerate innovation and fuel business growth.

DEFINING THE PROBLEM



Instacart faces a profit growth challenge due to suboptimal customer engagement and retention. Current personalization efforts underutilize rich transactional data, limiting the delivery of tailored shopping experiences and constraining revenue potential.

DATASET INFORMATION



kaggle

Available on Kaggle

3,421,083

Orders

Includes

- More than 200,000 users
- Order history
- Product catalog

DATA ANALYSIS METHODOLOGIES

Exploratory Analysis

Uncovered trends and insights, enabling a deeper understanding of transaction patterns.

Multi-level Association Rules

Revealed hidden relationships across product, aisle, and department levels, guiding targeted cross-selling and product recommendations.

Customer Segmentation

Identified distinct user groups based on purchasing behaviors, enabling targeted engagement strategies.

Feature Engineering

Developed targeted features to enable more precise and actionable predictive modeling.

Ensemble Optimization

Leveraged ensemble learning methods to foster model diversity and assess optimal performance options, complemented by sampling techniques to refine the dataset.

DATA
PRE-PROCESSING

PRE-PROCESSING

01

Feature Engineering

Mapped days of the week to categorical names and converted times to AM/PM format for enhanced interpretability.

02

Data Merging

Combined diverse data sources to create a unified dataset for comprehensive analysis.

03

Outlier Insights

Retained 9.17% outliers; refined mean validates the median while capturing unique customer behavior and trends.

OUTLIER REMOVAL

Summary with Outliers

Statistic	Amount
Count	206209
Mean	16.59
Std Dev	16.64
Min	4
25%	6
50%	10
75%	20
Max	100

Summary without Outliers

Statistic	Amount
Count	189192
Mean	12.57
Std Dev	8.90
Min	4
25%	6
50%	9
75%	17
Max	41

Removing outliers narrowed the data distribution and refined the mean. The median remained largely unchanged ,providing strong evidence that it is the most reliable measure of central tendency.

Our primary objective is to determine which previously purchased products are likely to appear in future customer orders, while also identifying trends and uncovering key patterns. Although 9.17% of our orders qualify as anomalies, we've decided to retain them. These outliers could reveal unique customer behaviors, niche preferences, or valuable insights that might otherwise be missed.

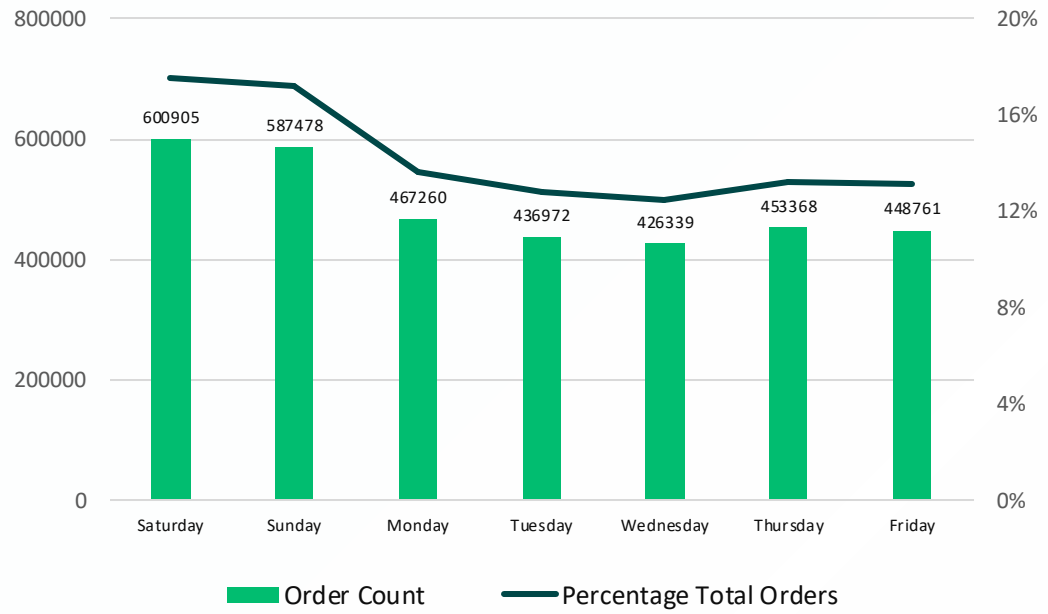
DATA
VISUALIZATION

ORDER METRICS SUMMARY

3,421,083	Number of Orders
8	Median Order Size
10	Median Number of Orders per User
453,368	Median Quantity of Orders per Day
13 Days	Median Time Between Orders

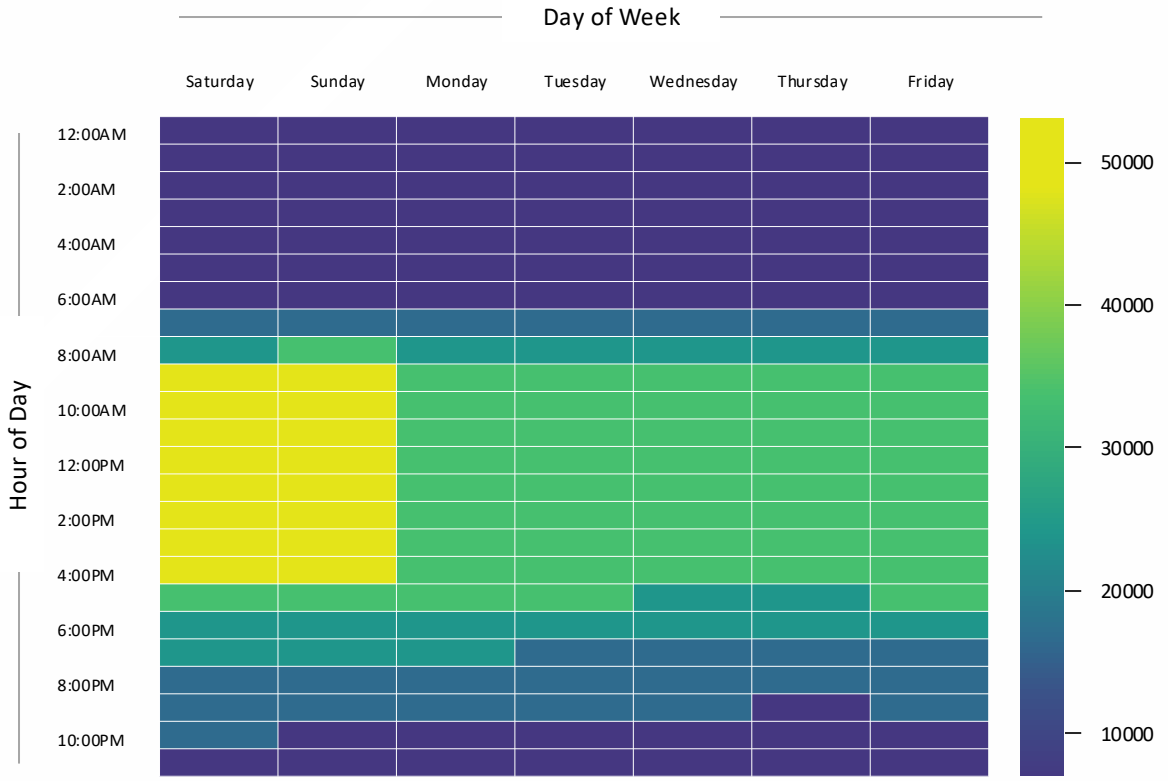
ORDER METRICS KEY FINDINGS

Orders by Day of Week (Percentage of Total)



35% of orders occur during the weekend

Consistent activity during the week



Peak hours are 10AM to 5PM

64% of orders occur during peak hours

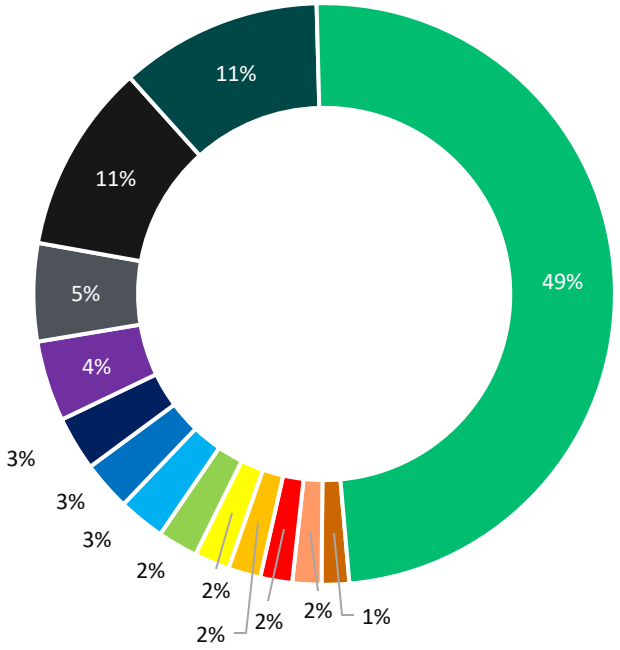
PEAK ACTIVITY
DAILY ANALYSIS

Day of Week	Percent of total orders during peak hours	Percent of daily orders during peak hours
Saturday	12	67
Sunday	11	64
Monday	9	63
Tuesday	8	63
Wednesday	8	62
Thursday	8	64
Friday	8	65

Each day consistently sees over 60% of its orders during periods of peak activity.

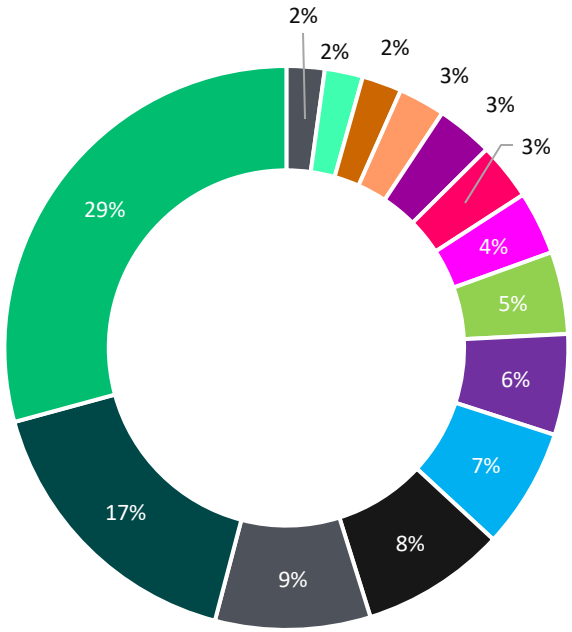
Peak hours contribute 12% of total orders on Saturday, 11% on Sunday, and between 8% and 9% on weekdays.

ORDER DISTRIBUTION BY AISLE AND DEPARTMENT



Aisle Distribution

1%	Ice Cream Ice
2%	Frozen Produce
2%	Refrigerated
2%	Bread
2%	Soy Lactosefree
2%	Chips Pretzels
3%	Sparkling Water
3%	Milk
3%	Packaged Cheese
4%	Yogurt
5%	Vegetables Fruits
11%	Fresh Vegetables
11%	Fresh Fruits
49%	Other

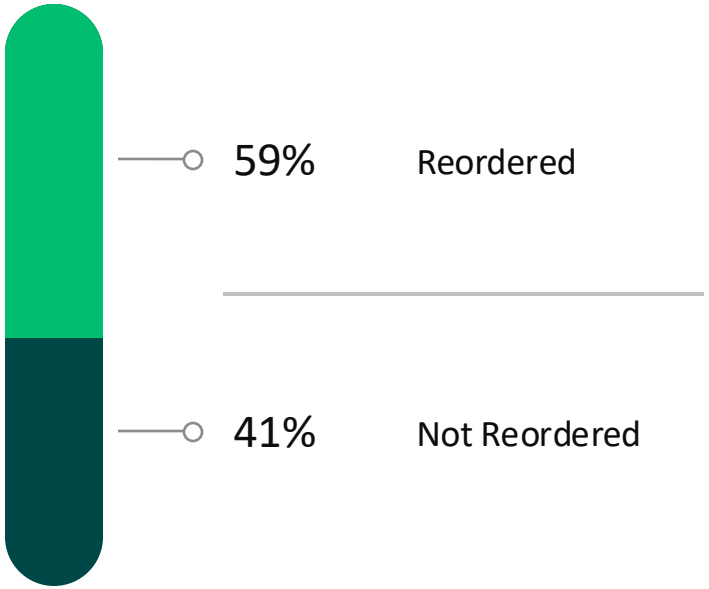


Department Distribution

2%	Produce
2%	Dairy Eggs
2%	Snacks
3%	Beverages
3%	Frozen
3%	Pantry
4%	Bakery
5%	Canned Goods
6%	Deli
7%	Dry Goods Pasta
8%	Household
9%	Breakfast
17%	Meat Seafood
29%	Other

REORDERED PRODUCTS

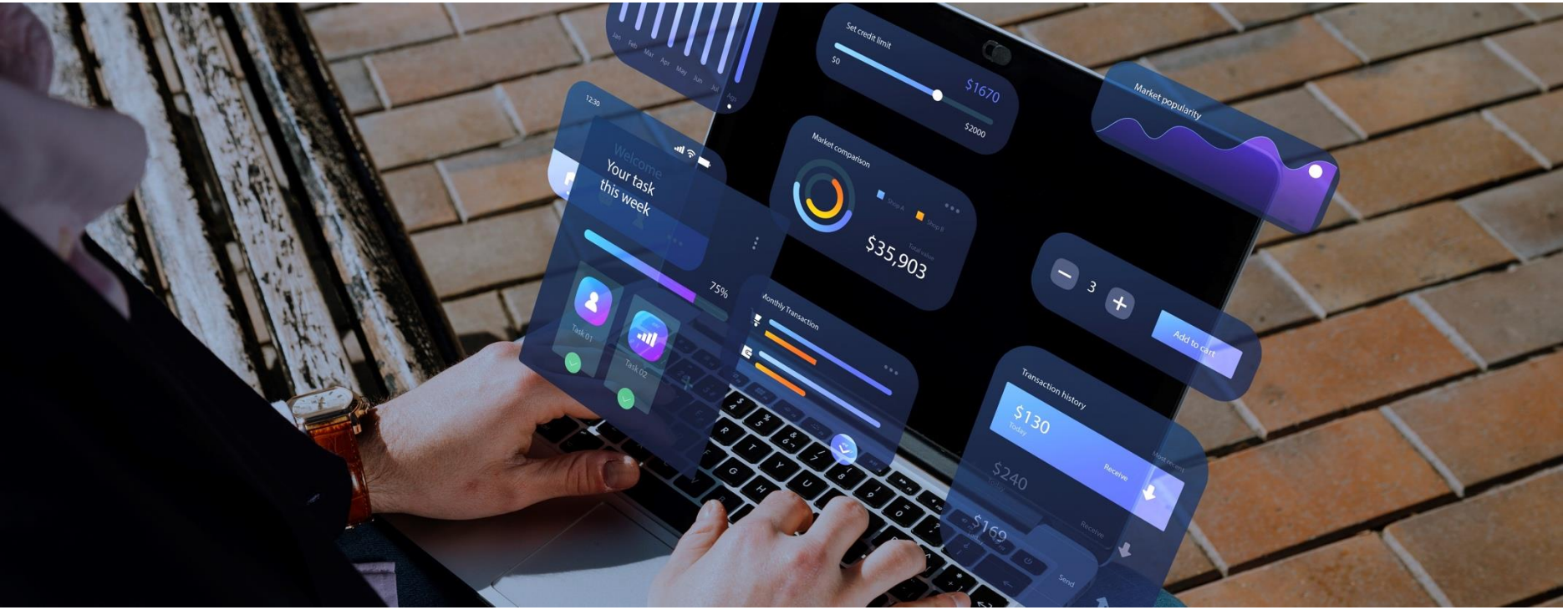
Reordered Products Distribution



Top 10 Reordered Products	Percent of total reorders
Banana	2.08
Bag of Organic Bananas	1.65
Organic Strawberries	1.08
Organic Baby Spinach	0.98
Organic Hass Avocado	0.89
Organic Avocado	0.70
Organic Whole Milk	0.60
Large Lemon	0.56
Organic Raspberries	0.55
Strawberries	0.52

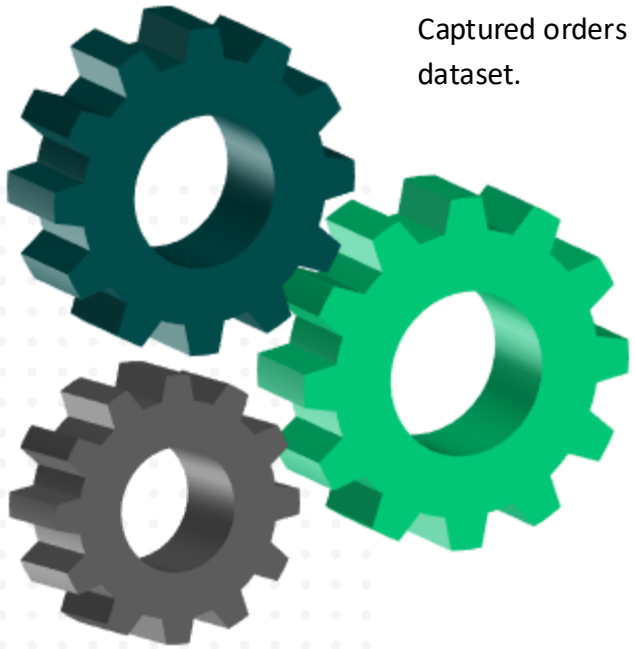
Organic products dominate the Top 10

Top 10 accounts for 9.61% total reorders


Set credit limit
$50
$1670
$2000
Market popularity
Market comparison
$35,903
Welcome
Your task
this week
75%
Task 01
Task 02
Monthly Transaction
3
Add to cart
Transaction history
$130
Today
Receive
$240
$169
Send

DATA MINING

DATA MINING METHODOLOGIES



Focused Timeframes for Impact

Captured orders during peak hours to ensure analysis focuses on the most active, high-impact periods. Achieved a 36% reduction in the overall dataset.

Targeted High-Value Customers

Isolated high-frequency customers during peak times to prioritize those with the greatest influence on revenue. Further refined the dataset by 75%

Representative Product Sample

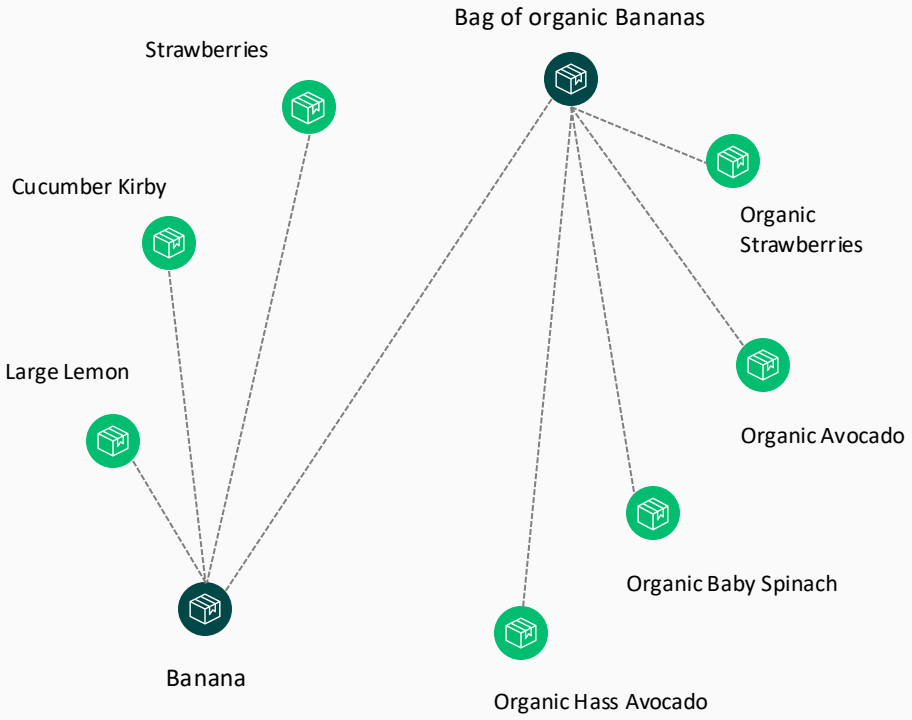
Focused analysis on the top 100 products to create a robust, yet manageable sample.

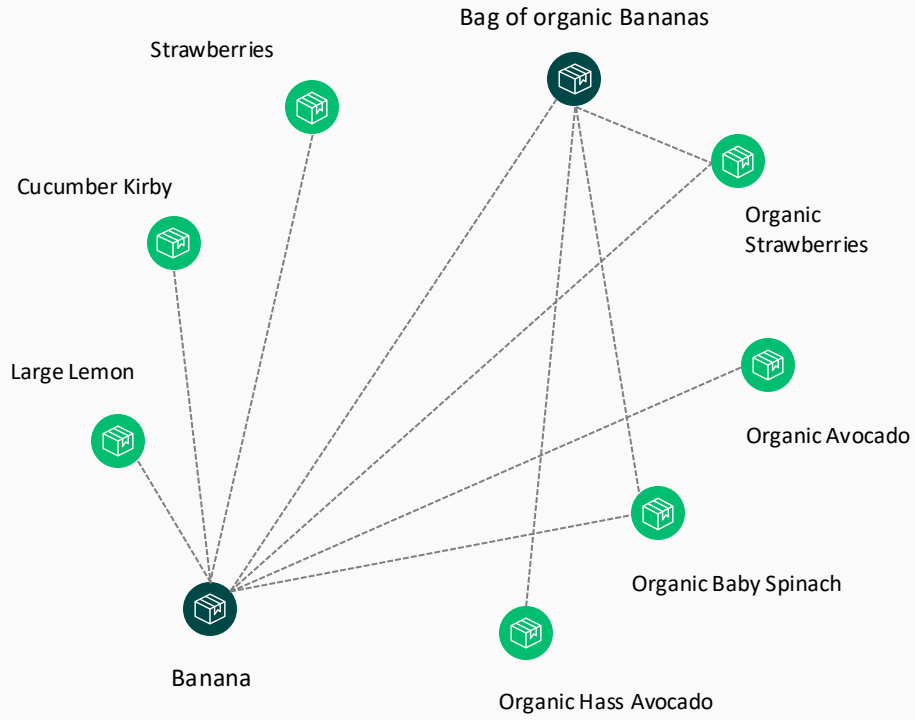
Strategic Outcome

Targeted approach provided actionable insights, enabling precise marketing and resource allocation.

PRODUCT ASSOCIATIONS

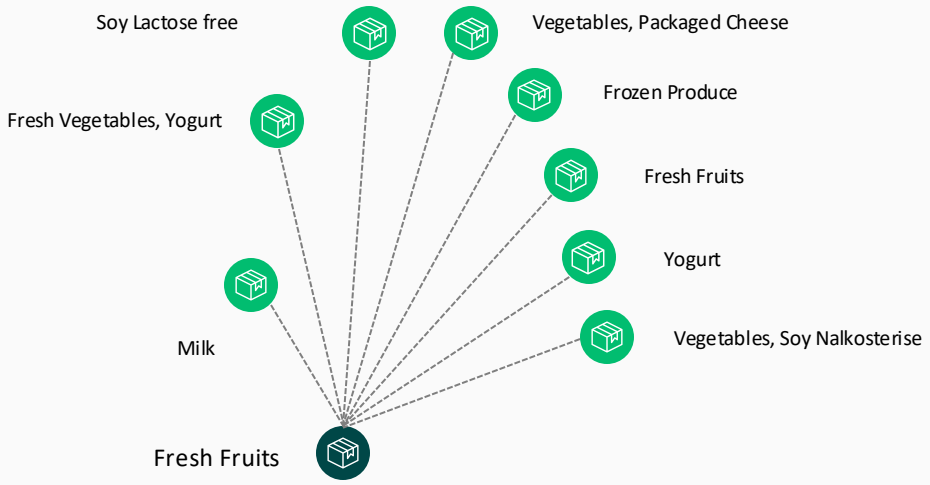
Key Node	Degree Centrality	Number of Strong Associations
Bananas	1.5	6
Bag of Organic Bananas	0.75	3

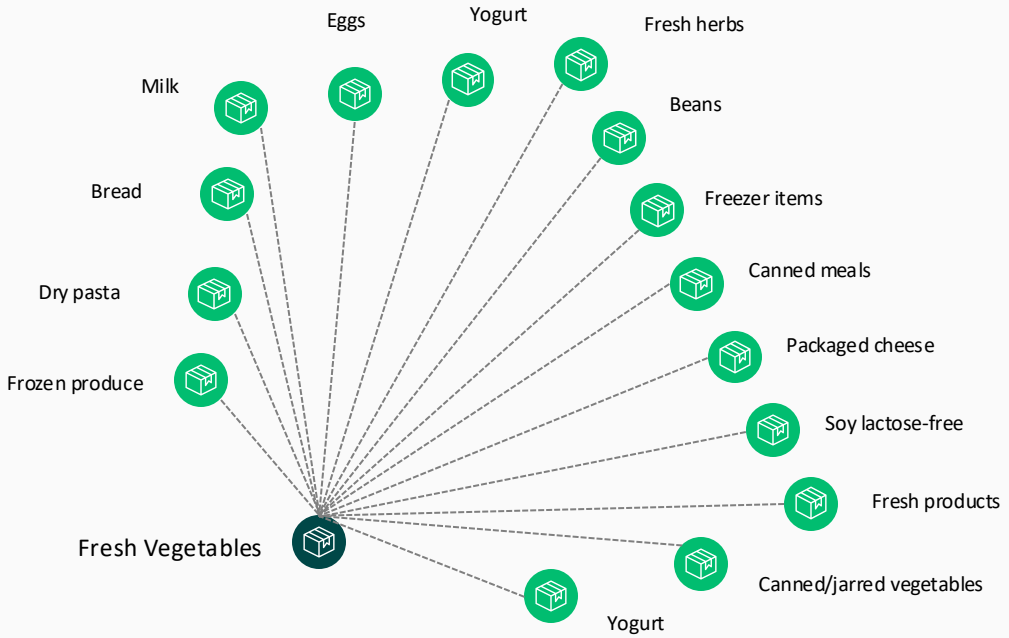




AISLE ASSOCIATIONS

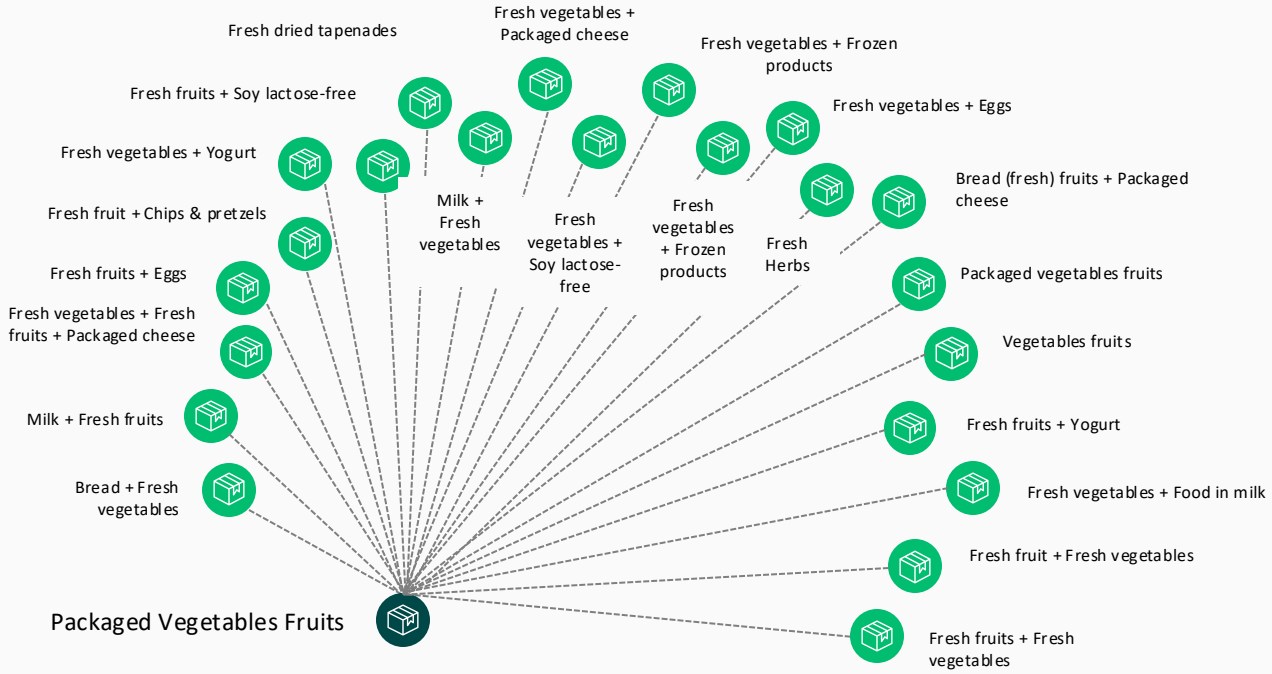
Key Node	Degree Centrality	Number of Strong Associations	Most Frequent Individual Products
Fresh fruits	0.21	9	Fresh & packaged vegetables, Milk, Soy lactose free, Yogurt, Frozen produce, Packaged cheese
Fresh vegetables	0.48	20	Packaged vegetables, Fresh fruits, Frozen produce, Eggs, Milk, Yogurt

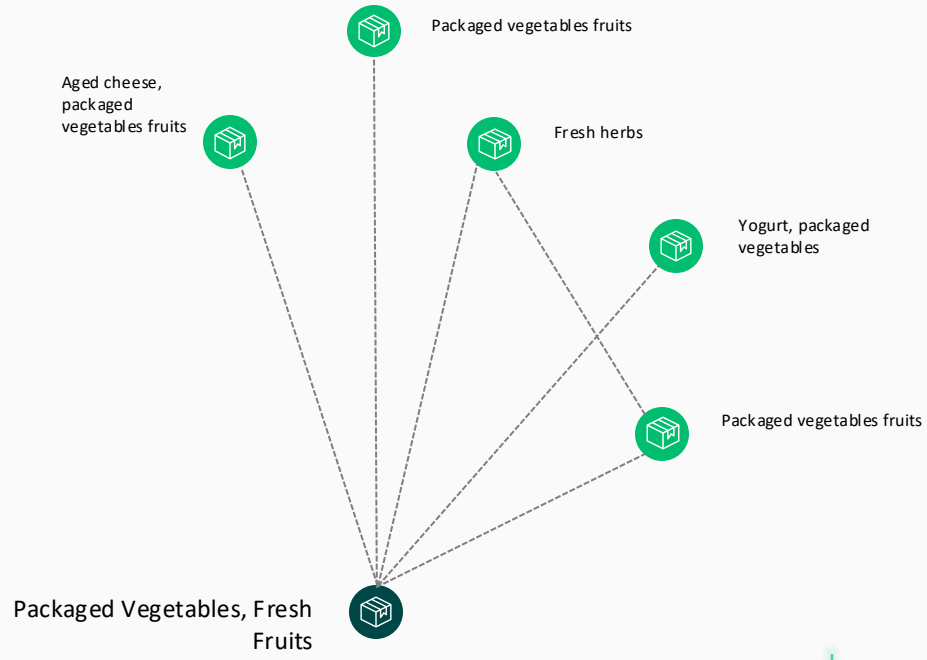




AISLE ASSOCIATIONS

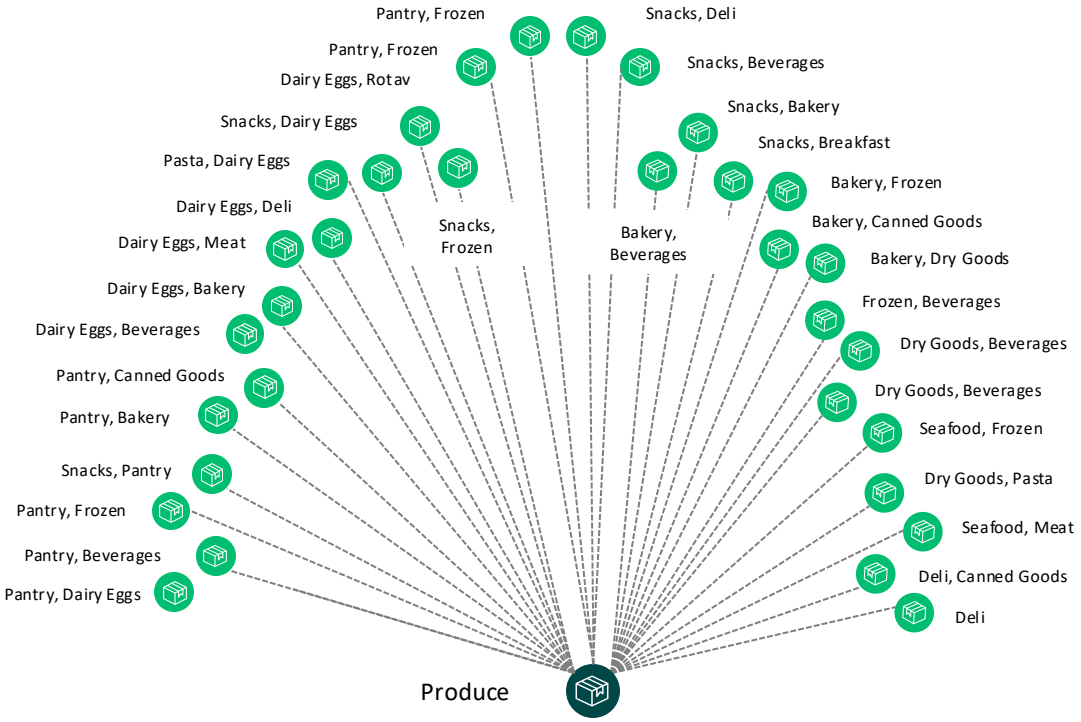
Key Node	Degree Centrality	Number of Strong Associations	Most Frequent Individual Products
Packaged Vegetables, Fruits	0.55	22	Fresh fruits, Fresh vegetables, Frozen produce, Eggs, Packaged cheese, Milk
Fresh Vegetables, Fresh Fruits	0.14	5	Packaged vegetables, Fresh fruits

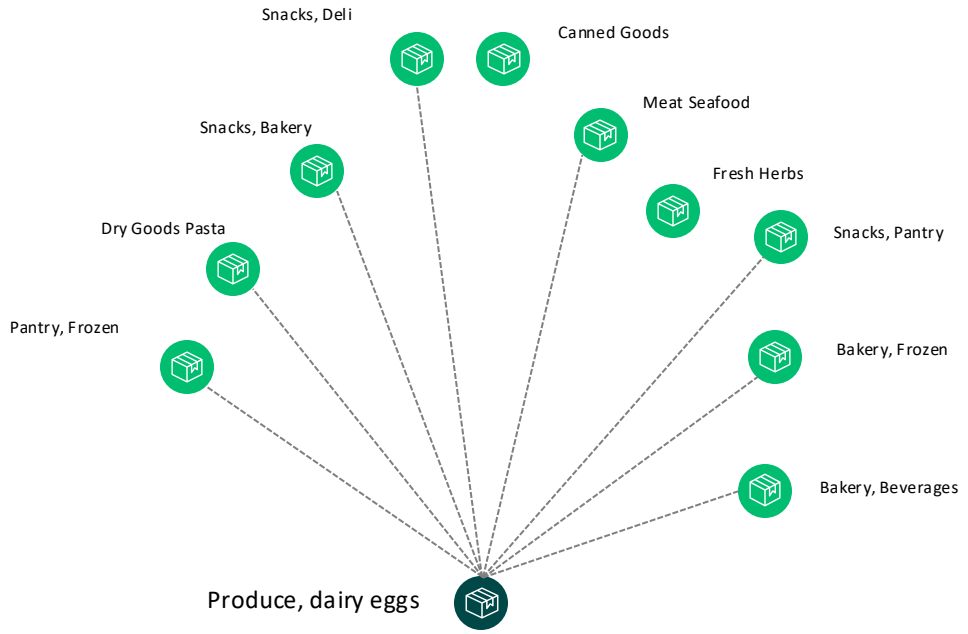




DEPARTMENT ASSOCIATIONS

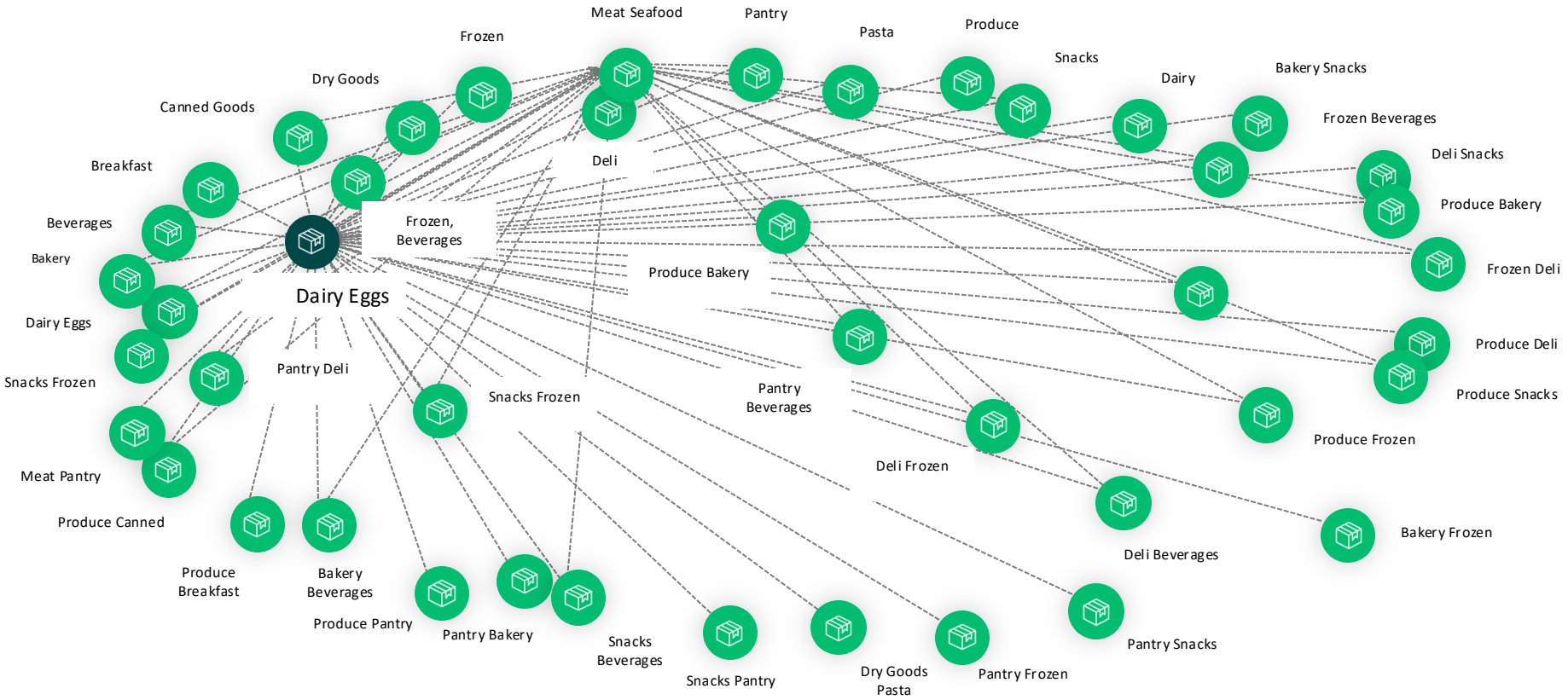
Key Node	Degree Centrality	Number of Strong Associations	Most Frequent Individual Products
Produce	0.70	46	Dairy Eggs, Frozen, Pantry, Beverages, Meat, Pantry
Produce, dairy eggs	0.15	10	Snacks, Pantry, Frozen





DEPARTMENT ASSOCIATIONS

Key Node	Degree Centrality	Number of Strong Associations	Most Frequent Individual Products
Dairy eggs	0.66	42	Frozen, Produce, Pantry, Snacks, Pantry , Beverages



Product-Level Insights

Strategic Anchor Products

Banana | is highlighted as a pivotal product. Its prominence indicates it is a frequent co-purchase partner with many other items.

Opportunity | Leverage bananas for targeted promotions, cross-selling, and prime shelf placement to drive overall basket growth.

Aisle-Level Dynamics

Integrated Fresh & Packaged Produce

Fresh | fruits and vegetables both show strong linkages with complementary items such as packaged vegetables, dairy, and frozen products.

Packaged vegetables, fruits stands out with the highest aisle-level centrality (0.55) and a robust set of 22 strong associations.

Opportunity | Consider bundled offers or thematic displays that group fresh and packaged produce with related dairy and frozen items.
Optimize aisle layout to place frequently co-purchased items in closer proximity to enhance convenience and stimulate incremental sales.

Smaller Nodes as Niche Opportunities

The Fresh vegetables, fresh fruits node, while less central (0.14), indicates targeted niches where curated promotions could drive attention and trial.

Department-Level Insights

Produce as a Sales Engine

The standalone Produce department exhibits the highest centrality (0.70) with 46 strong associations, underlining its role as the primary driver in shopping baskets.

Opportunity

Focus on strategic placement and marketing within the produce section to leverage its broad influence.

Dairy and Eggs – A Complementary Powerhouse

Dairy eggs also show high centrality (0.66) and robust cross-category associations with frozen, pantry, and snacks.

Opportunity

Implement cross-promotional strategies that pair dairy and eggs with these categories, encouraging a diverse basket and enhancing customer satisfaction.

Combined Category Nuances

The merged Produce, dairy eggs node, with a lower centrality (0.15), suggests that these departments, while important on their own, may perform better when their strategies are tailored individually rather than combined.

Strategic Recommendations

Promotional Strategies

Develop targeted promotions around high-centrality products like Bananas, and design bundled offers that leverage the co-purchase behaviors seen in produce and dairy eggs.

Integrated Marketing Campaigns

Craft campaigns that emphasize the complementary nature of items across fresh, packaged, and dairy categories, encouraging customers to explore a wider range of products in a single trip.

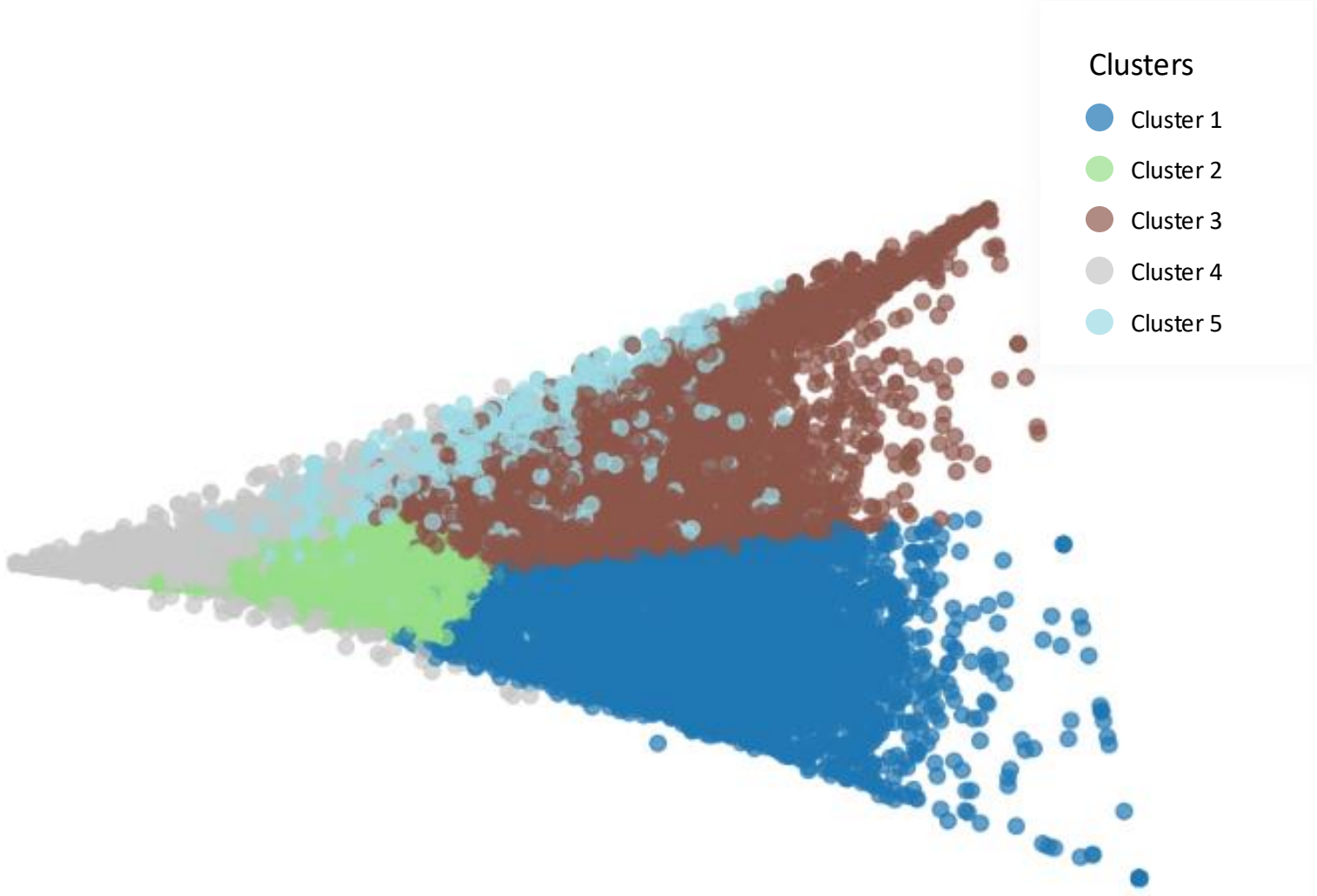
Store Layout Optimization

Reconfigure store layouts to place items with strong co-purchase links in closer proximity. This could improve convenience and spur additional purchases.

Inventory & Supply Chain Management

Prioritize stocking and inventory management in high centrality areas to minimize stockouts and support the increased demand driven by these associations.

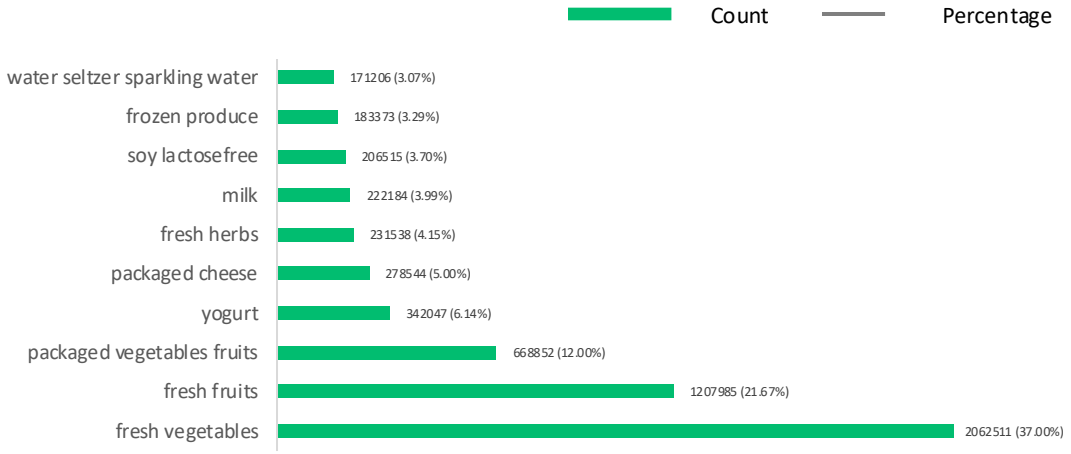
Customer Segmentation



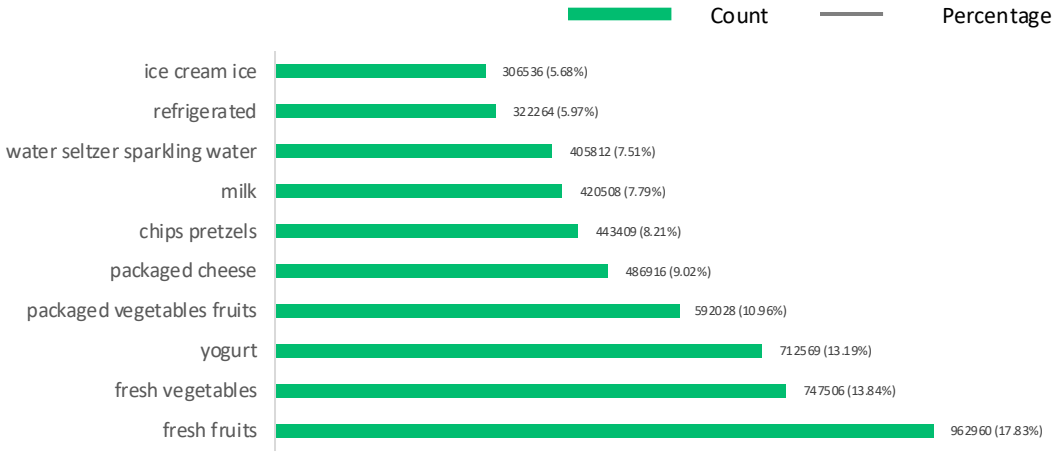
We used the elbow method and silhouette scores to validate our data-driven clustering, confirming five distinct groups.

These five segments offer clear insights into customer behavior, enabling targeted strategies and improved decision-making.

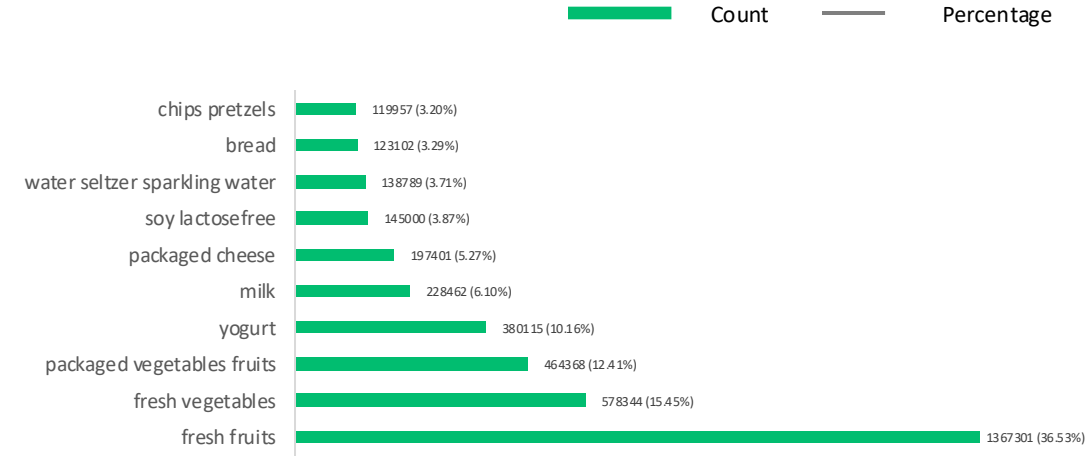
Cluster 1
56k customers (27%)



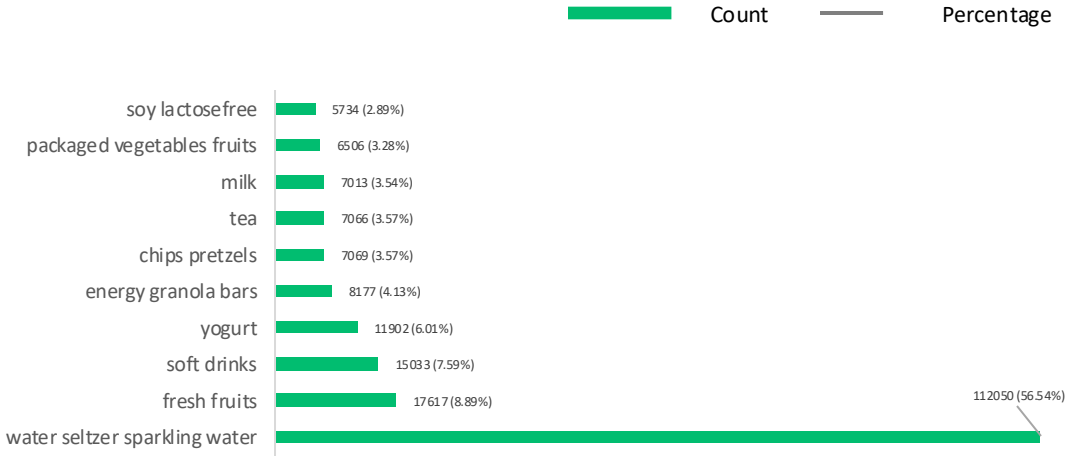
Cluster 2
99k customers (48%)



Cluster 3
38k customers (19%)

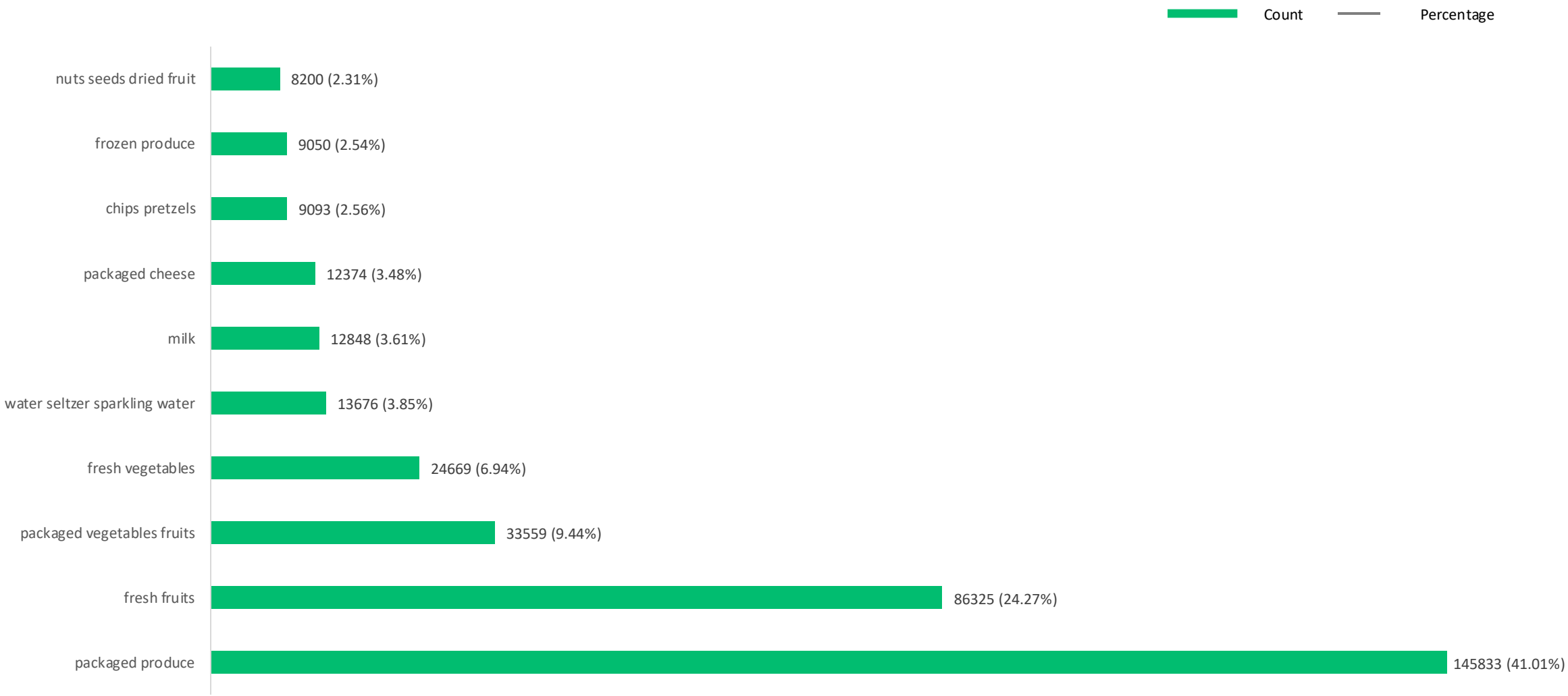


Cluster4
5k customers 2.6%



Cluster 5

8k customers (3.9%)



Customer Segmentation Insights

Segmenting at the aisle level strikes the optimal balance: it's granular enough to uncover consistent co-purchase patterns, yet broad enough to avoid the noise and volatility of individual SKUs. This level of analysis reveals meaningful behavioral clusters that inform both personalization and merchandising strategies.

Cluster	Segment Name	Size	Description
1	Health-Conscious Staples	27%	Prioritizes fresh produce, yogurt, soy/lactose-free items, and packaged vegetables. Indicates a wellness-driven, ingredient-aware customer base receptive to clean-label and organic promotions.
2	Balanced and Indulgent	48%	Largest segment with both core staples (milk, yogurt, fresh produce) and indulgent/snack items (ice cream, chips). Ideal for broad campaigns that combine health and treat-based messaging.
3	Core Staples	19%	Leans on bread, milk, cheese, yogurt, and fresh produce. Represents a conventional, family-focused basket—well suited for promotions around household staples and weekly meal planning.
4	Beverage-Centric Niche	2.6%	Dominated by sparkling water, with moderate interest in tea, soft drinks, and snacks. Indicates a small but highly focused group ideal for beverage brand partnerships and loyalty bundles.
5	Packaged Convenience Shoppers	3.9%	Strong preference for packaged produce, fresh fruits, and nuts/dried fruit. Suggests grab-and-go and snack pack appeal—ideal for ready-to-eat promotions and time-saving meal solutions.

Fresh Produce Anchors Loyalty

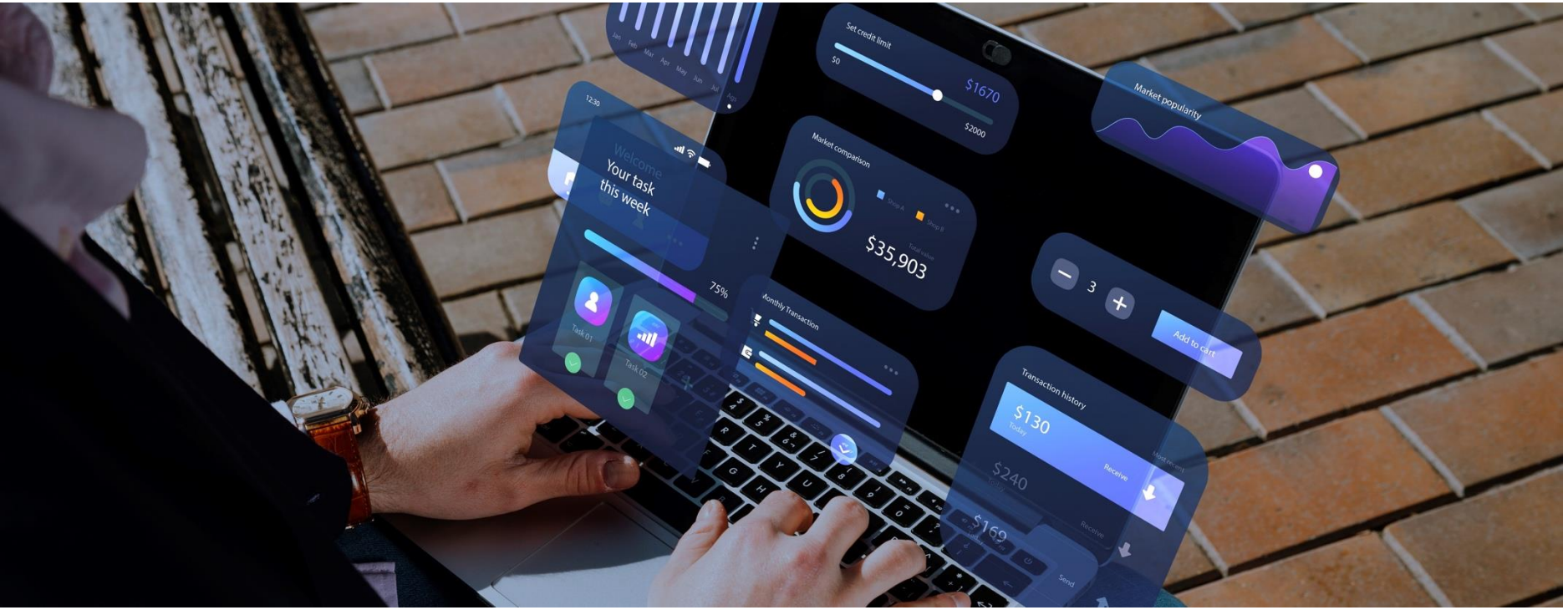
Fresh fruits and vegetables are consistently top-ranked across all clusters, making them a strategic cornerstone for retention and recommendation systems.

Cross-Selling Opportunity

Frequent co-purchases in dairy, produce, and snacks highlight opportunities for targeted bundles (e.g., yogurt + fresh fruit, milk + cereal alternatives).

Strategic Takeaway

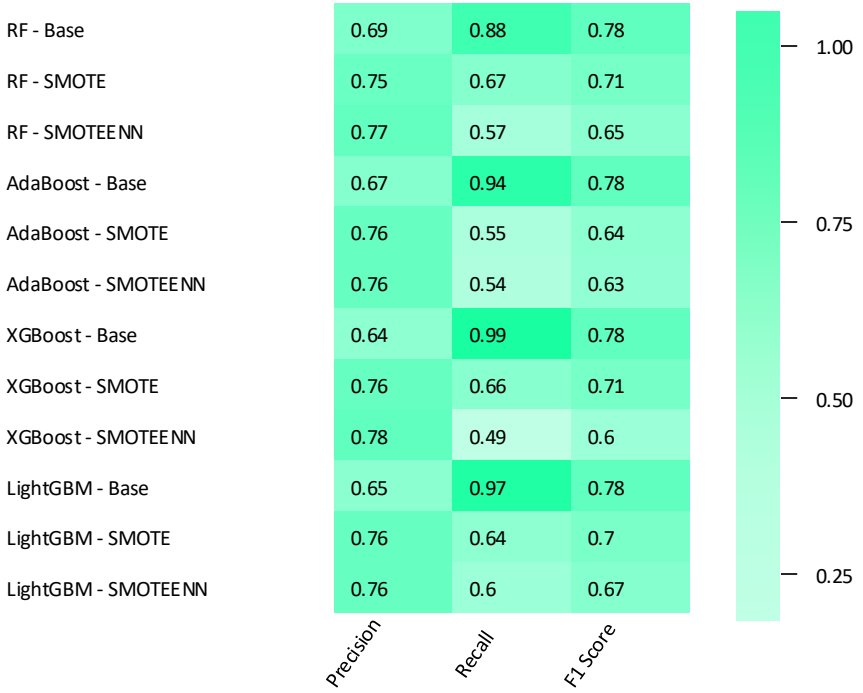
Use fresh produce as a universal hook across segments, while tailoring upsells and messaging based on segment-specific priorities.



DATA MODELING

Model Comparison (Metrics)

Classification Report Heatmap (All Models)



Confusion Matrix Heatmap (All Models)

Chose XGBoost + SMOTE for its superior balance between capturing true reorders and reducing false positives; we'll next fine-tune its hyperparameters and decision threshold to maximize ROI and operational efficiency.

Best Model XGBOOST SMOTE TUNED

Captures genuine reorder signals while slashing false alarms—aligning inventory and promotional spend with true demand to maximize ROI.

Class	Precision	Recall	F1-Score	Support
0	0.58	0.51	0.54	11229885
1	0.73	0.78	0.76	19126536
Accuracy			0.68	30356421
Macro Avg	0.65	0.65	0.65	30356421
Weighted Avg	0.67	0.68	0.68	30356421

TN	FP	FN	TP
5711005	5518880	4167482	14959054

True Negatives

Correctly predicted 5.7M non-reorders

False Negatives

Missed 4.16M reorders

True Positives

Correctly predicted 14.95M reorders

False Positives

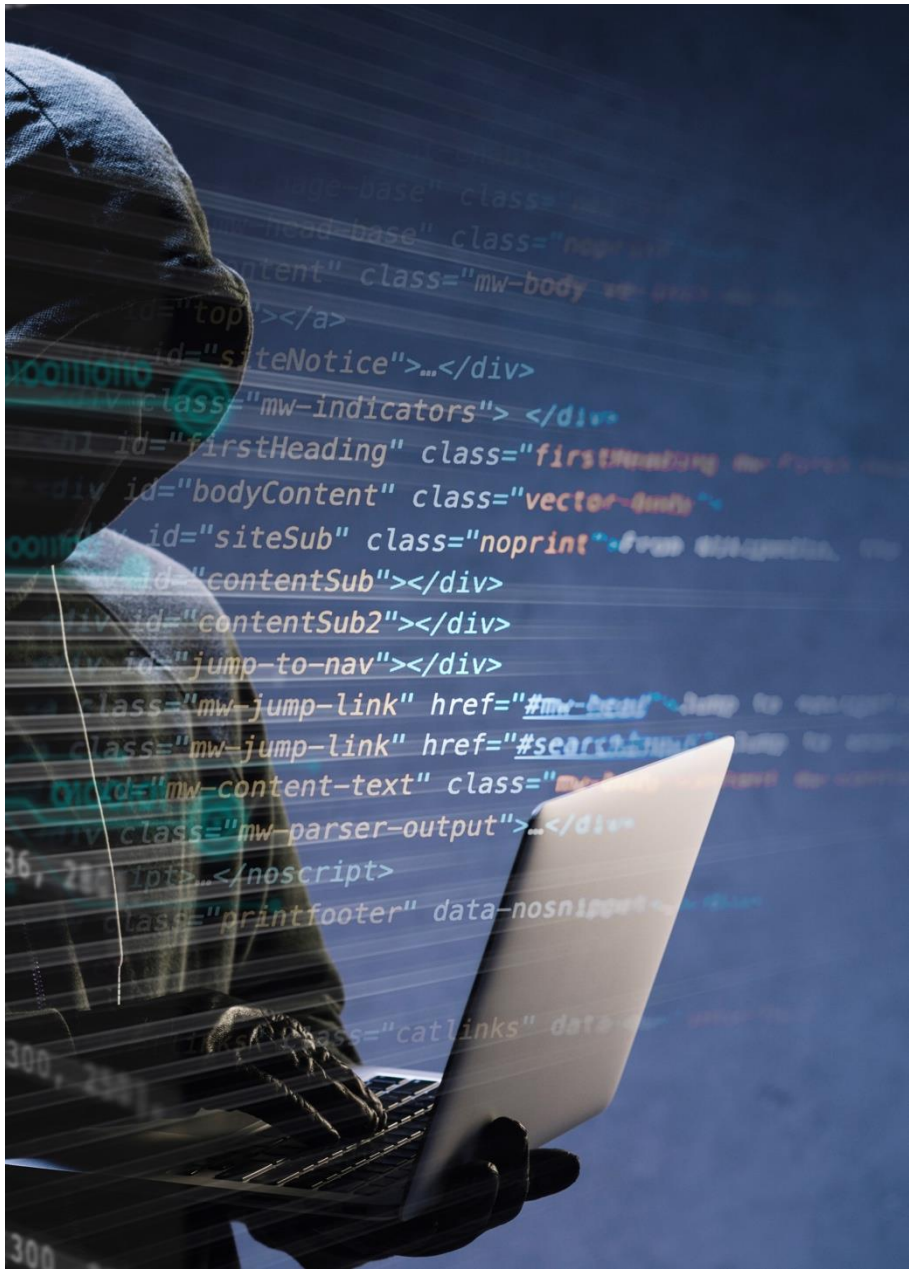
Mispredicted 5.5M reorders

Positive class

High recall (0.78), high precision (0.73)
Captures 78% reorders with 73% precision

Negative class

Mediocre performance with recall (.51) and precision (.58)



MODEL CONCLUSIONS

The optimized XGBoost model with SMOTE captures 78% of reorder predictions with 73% precision, outperforming all models.

Other models, while demonstrating strong capture rates, skew heavily toward false positives—driving up promotion and inventory costs and eroding ROI:

- LGBM Base: Captures 97% of fraud with 65% precision.
- AdaBoost Base: Captures 94% of fraud with 67% precision.
- Random Forest Base: Captures 88% of fraud with 69% precision.

CONCLUSION

Our Instacart reorder model, trained on millions of transactions, captures 78% of reorders with 73% precision while keeping noise to a minimum. Embedding these insights into real time recommendations drives repeat purchases, deepens customer engagement and unlocks sustainable revenue growth.

ANY QUESTIONS?

THANK YOU FOR YOUR TIME

Charles Bryant