MACHINE LEARNING & AI AT SCALE

Data Understanding Assignment

Prepared By:
Zach Wei
Xavier Liu
Cynthia Kong
Ivanie Stella Umuhoza
Sandrine Bakuramutsa

MARCH 8, 2025

GOIZUETA BUSINESS SCHOOL
MSBA

Table of Contents

Exploratory Data Insights Report

1 Project Overview

1.1 Background

ACSE Supermarket is a large retail chain operating over 40 stores across North America. With a product catalog exceeding 100,000 items across more than 100 categories, the company serves a vast customer base, including both regular shoppers and members of the ACSE Rewards program, which provides access to exclusive promotions and discounts.

To optimize supply chain management, store operations, supplier relations, pricing, promotions, and marketing, ACSE aims to implement a recommender system. This system will help the company make data-driven decisions on product assortment, shelf space allocation, promotions, reorder levels, and purchasing patterns. By leveraging transaction history and product data, ACSE seeks to enhance customer experience, maximize revenue, and improve operational efficiency.

1.2 Problem Statement

ACSE Supermarket faces increasing competition in the retail grocery sector and needs to enhance customer experience, maximize revenue, and improve operational efficiency. The company lacks insights into customer purchasing patterns, product performance, and store operations needed to make data-driven decisions. Without a systematic approach to analyzing their vast transaction data, ACSE is missing opportunities to optimize product assortment, shelf space allocation, promotions, reorder levels, and purchasing patterns.

1.3 Objectives

This exploratory data analysis aims to analyze customer purchasing behavior to identify distinct segments and create opportunities for targeted marketing strategies. We will evaluate product performance across various categories to optimize inventory management and shelf space allocation in stores. The analysis will identify top-performing and underperforming stores to develop location-specific strategies that address each store's unique challenges and opportunities. Additionally, we will discover frequently purchased product combinations to inform cross-selling opportunities and bundle promotions. Ultimately, this analysis will develop a foundation for a data-driven recommender system that will enhance customer experience and drive sales growth across the supermarket chain.

1.4 Available data

The analysis is based on comprehensive transactional data from ACSE Supermarket, providing a rich source of customer insights and product performance metrics. This dataset includes detailed customer purchase history capturing shopping patterns over time, extensive product information with categories, descriptions, and product types that allow for hierarchical analysis, and complete transaction details including dates, quantities, and revenue figures. The data also encompasses store information enabling location-based analysis and performance comparisons, as well as promotional data that provides visibility into discount effectiveness and customer response to various marketing initiatives. This multi-dimensional dataset enables a thorough exploration of the relationships between customers, products, and store operations.

2 Business Understanding

2.1 Industry Trends in Retail Analytics

The retail grocery industry is experiencing significant transformation. ACSE Supermarket is aiming to implement a recommender system to optimize supply chain management, store operations, supplier relations, pricing, promotions, and marketing. This system will help the company make data-driven decisions on product assortment, shelf space allocation, promotions, reorder levels, and purchasing patterns.

Implementing data-driven approaches is becoming essential in the retail industry, with advanced retailers leveraging analytics to optimize everything from inventory management to customer experience.

2.2 Business Benefits of Customer and Product Analysis

Analyzing customer behavior and product performance offers ACSE Supermarket several potential benefits:

By leveraging transaction history and product data, ACSE seeks to enhance customer experience, maximize revenue, and improve operational efficiency. The analysis will help identify high-value customers and their preferences, allowing ACSE to develop targeted loyalty programs.
Understanding which products drive revenue and customer traffic will enable ACSE to allocate shelf space more effectively and ensure that high-demand items are consistently available. The analysis of category performance and promotion effectiveness will help optimize inventory and marketing strategies.

Market basket analysis will reveal product combinations that are frequently purchased together, enabling cross-selling opportunities and improved store layouts. These insights will allow ACSE to create compelling product bundles and suggest complementary items to customers.

2.3 Key Questions Guiding the Analysis

This exploration aims to answer critical business questions including:

- What are the products and product groups with the best volumes, revenues, profits, transactions, and customers?
- Are there product groupings beyond traditional categories, such as Key Value Items (KVI) and Key Value Categories (KVC), traffic drivers, and promotion-based classifications?
- How do customers segment based on purchasing behavior, including spending levels, shopping frequency, product variety, and price sensitivity?
- How do stores perform relative to each other, and how can they be segmented to develop targeted strategies?
- What product combinations are frequently purchased together, and how can this inform merchandising and promotion strategies?

The analysis of these questions will provide actionable insights to enhance ACSE's competitive position in the retail market.

3 Data Understanding and Descriptive Data Analysis

3.1 Overview of Data Sources and Sampling Methodology

The analysis is based on comprehensive transactional data from ACSE Supermarket. To ensure representative results across the entire store network, we employed a stratified sampling approach by store. This methodical sampling technique guaranteed that data from all retail locations was proportionally represented in the analysis, preventing any bias that might occur from overrepresentation of high-volume or underrepresentation of low-volume stores.

The dataset consists of 1,206,131 original transactions, which were all preserved throughout the cleaning process. Rather than removing anomalies, the approach focused on adding contextual flags to enhance analytical potential while maintaining data integrity.

Key variables in the analysis include:

- Transaction data: trans_id, store_id, cust_id, trans_dt, sales_qty, sales_amt
- Product data: prod_id, prod_desc, prod_category, prod_subcategory, prod_type
- Customer metrics: cust_id,total_purchases, total_stores_visited, total_transactions,unique_products_purchased,total_refunds,net_revenue,refund_ratio, spending_category,shopping_frequency_category,product_variety_category,avg_transaction_value,pricing_behavior_category,dominant_product_category
- Store metrics: store_id, total_transactions, unique_customers,total revenue, customer count, total_sales_volume, avg_basket_size, avg_revenue_per_transaction, avg_purchase_frequency, store_cluster

3.2 Data Quality Assessment

3.2.1 Missing Values Analysis

The analysis of missing values involved a thorough examination of both the transactions and products datasets to identify any gaps that could compromise data quality or analytical results. This process was approached systematically after performing stratified sampling by store to ensure representative data across all retail locations. The analysis began by loading the stratified sample data from the CSV files and using pandas' built-in functionality to detect null values across all columns, then creating a comprehensive summary of missing values by column to quantify the extent of any data gaps, and finally analyzing the potential impact of missing values on subsequent analysis and the recommender system development.

The results showed strong data integrity in the transactions dataset with zero missing values across all transaction records, indicating robust point-of-sale data collection systems. However, in the products dataset, 2,240 missing values were discovered specifically in the prod_type field. This pattern of missing values concentrated in a single field suggests a systematic issue with product type classification rather than random data collection errors. While the transaction data showed complete coverage, having these missing product types could potentially impact the granularity of product recommendations.

This finding of mostly complete data with specific, concentrated gaps provides valuable guidance for the recommender system project. The completeness of transaction records means that customer purchase histories are intact, providing reliable behavioral patterns for recommendation algorithms. The missing product types, while representing a data quality issue, are contained to a single field and can be managed through consistent placeholder values. This allows the team to proceed with high confidence in the transaction data integrity while implementing targeted strategies to account for the missing product classifications when developing category-based recommendations.

3.2.2 Data Inconsistencies

The analysis of inconsistent product descriptions and category mappings involved identifying formatting inconsistencies that could impede accurate product categorization and subsequent recommendation quality. The code implemented a targeted text standardization approach focusing on three key product attributes. For product descriptions, the system converted all text to lowercase and removed leading and trailing whitespace using standard string transformation functions (strip and lowercase). This standardization ensures consistent text formatting for accurate matching and analysis, eliminating variations due to capitalization and whitespace that could create artificial distinctions between products. Similarly, for product categories and subcategories, the process applied whitespace trimming to ensure consistent formatting for correct grouping in analyses and to prevent duplicate categories caused solely by whitespace variations.

For handling logical inconsistencies, the approach focused on contradictions in the data, particularly transactions with zero or negative quantities but positive sales amounts. The

analysis identified 57 such transactions, which were corrected by setting the quantity to 1 while preserving the sales amount. This correction strategy maintained the transaction's financial impact while resolving the logical impossibility of selling a negative or zero quantity of an item. Another significant inconsistency addressed was the handling of negative sales values, which represented valid return transactions rather than errors. These 15,321 transactions (1.27% of the dataset) were preserved and flagged rather than removed, maintaining important business information about return patterns while ensuring they wouldn't artificially deflate sales metrics in subsequent analyses.

3.2.3 Outlier Detection

The analysis of outliers in customer purchases and store transactions aimed to identify unusual patterns that could represent either valuable business segments or potential data anomalies. Using statistical methods, the team systematically detected multiple types of outliers in the dataset:

- Customer spending: The Z-score method with a threshold of 3 standard deviations identified 3,108 customers with abnormally high total expenditures, with the most extreme spending reaching $10,042.44 compared to a threshold of $75.08.

- Transaction frequency: Similar analysis revealed 20,811 customers conducting an unusually high number of transactions, with some customers making up to 90 purchases during the analysis period.

- Product variety: 20,951 customers were flagged for purchasing an atypical variety of products, with some customers buying up to 88 different unique products compared to a typical range of fewer than 6 products.

- Store performance: Store-level analysis used a percentile-based approach to identify transaction count anomalies, revealing 3 stores with unusually low transaction volumes (fewer than 21 transactions).

Rather than removing these outliers, the approach flagged them in the dataset using boolean indicators ('is_outlier_customer' and 'is_low_activity_store'), thereby preserving all data while making anomalies easily identifiable for specialized analysis.

3.2.4 Final Cleaned Dataset

The final cleaned dataset represents the culmination of a comprehensive data cleaning process that prioritized data preservation while systematically addressing inconsistencies and anomalies. Through a series of targeted interventions, the cleaning process made significant improvements to the dataset while maintaining its fundamental integrity. All 1,206,131 original transactions were preserved, with no records being removed during cleaning. Instead, the approach added valuable contextual flags that enhance the analytical potential of the data.

The most substantial enhancement came through the identification and flagging of outlier patterns. By applying statistical methods, the process identified 23,458 unique customers with unusual purchasing behaviors across spending, transaction frequency, or product variety dimensions. Similarly, 3 stores with abnormally low transaction counts were flagged for targeted analysis. Text standardization efforts ensured consistent formatting across product descriptions and categories, eliminating variations due to capitalization and whitespace that could have created artificial distinctions.

The data integrity validation process confirmed that the cleaning interventions preserved all original business patterns while enhancing data quality. The cleaned dataset was successfully saved to CSV files, ready for further analysis and recommender system development. This conservative approach to data cleaning provides ACSE with maximum analytical flexibility while ensuring that unusual patterns are easily identifiable.

4. Key Insights & Findings

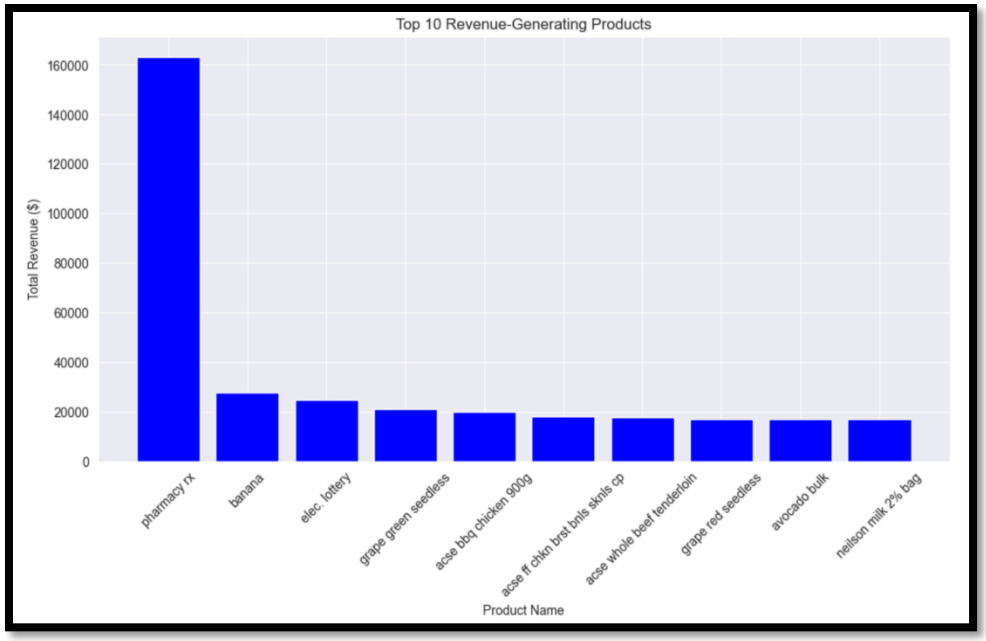
4.1 Product_Level Insights

4.1.1 Top Products and Categories by Revenue

The objective of this analysis was to identify the top-performing products based on key business metrics, including sales volume, revenue, transactions, and unique customers. By analyzing transaction data, we aimed to determine which products generate the highest revenue and have the most customer engagement. To evaluate product performance, we aggregated key metrics from the transaction data, including:

- Total Units Sold: The total quantity of each product purchased ("sales_qty")
- Total Revenue: The total sales amount before discounts ("sales_amt")
- Total Transactions: Number of distinct transactions in which the product appeared ("trans_id")
- Unique Customers: The number of different customers who purchased the product ("cust_id")

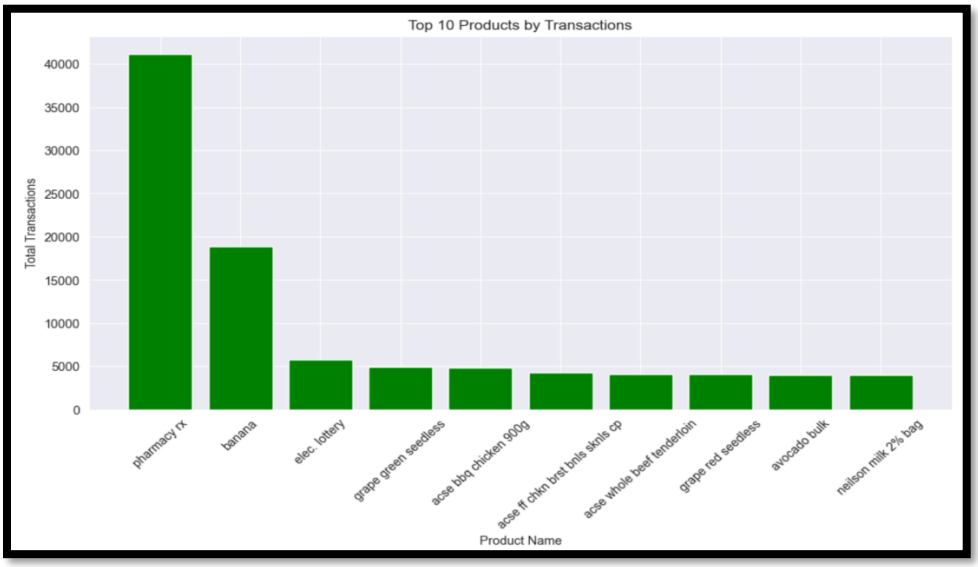
To gain deeper insights, the top 10 revenue-generating products were identified.



Key Findings:

- "Pharmacy Rx" is the leading revenue generator, significantly outperforming other products
- Other high-revenue items include bananas, electronic lottery, and fresh produce like grapes
- Premium cuts of meat (beef tenderloin, chicken breast) also contribute significantly to sales

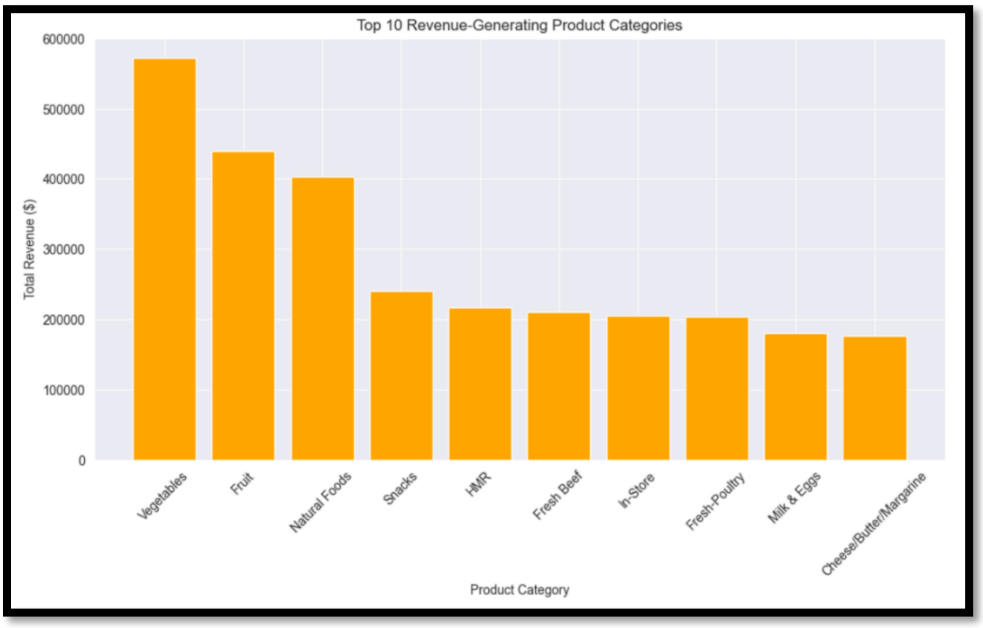
Revenue alone does not always indicate customer preference. To understand product popularity, we analyzed the top 10 products based on transaction count, which represents the frequency of purchases



Key Findings:

- "Pharmacy Rx" remains the most frequently purchased product
- Bananas have the second-highest transaction count, indicating strong customer demand for fresh produce
- Unlike revenue rankings, some lower-priced essential goods appear frequently in transactions

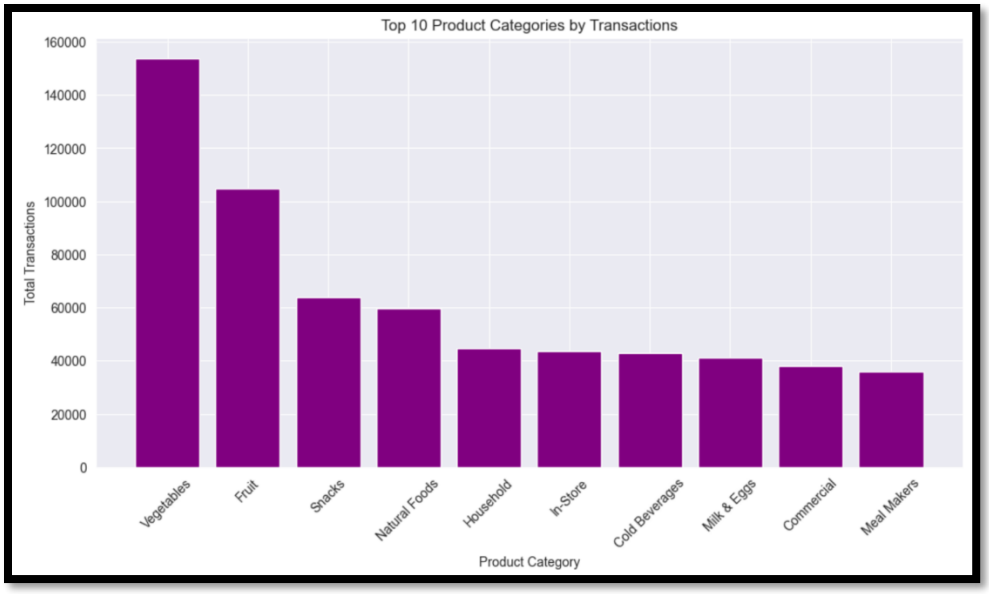
By analyzing total revenue across categories, we identified the highest-performing product groups.



Key Findings:

- Vegetables and Fruits generate the highest revenue, indicating strong consumer demand for fresh produce
- Natural Foods and Snacks follow closely, reflecting customer preference for healthier and convenient food options
- Protein categories (Fresh Beef, Poultry, and Dairy Products) contribute significantly to total revenue

Revenue alone does not fully capture customer preferences. Analyzing transaction count helps determine which categories are purchased most frequently.



Key Findings:

- Vegetables and Fruits remain the top categories, reinforcing their importance in customer baskets
- Snacks and Natural Foods appear frequently in transactions, suggesting they are staple purchases
- Household, Beverages, and Dairy Products also see high transaction volumes, highlighting essential everyday purchases

4.1.2 Promotion Analysis

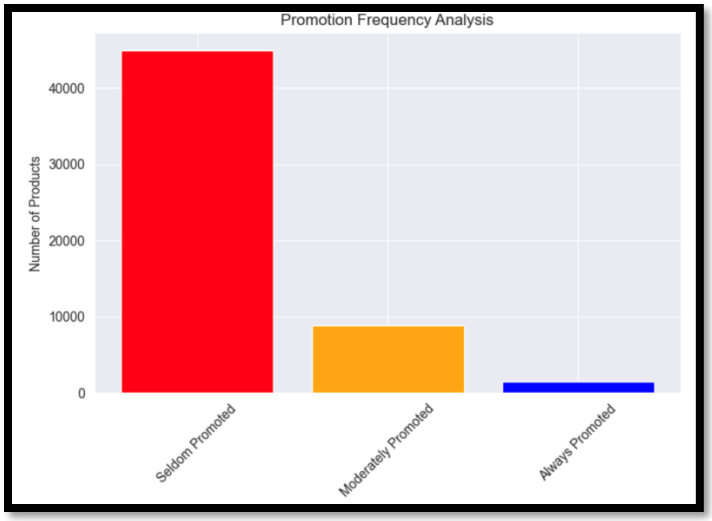
Beyond traditional product categories, we analyzed alternative product groupings that influence sales and customer behavior. This section focuses on:

- Promoted Products: Identifying products that are always, moderately, or seldom promoted based on pricing fluctuations

- Key Value Items (KVI) & Key Value Categories (KVC): Identifying high-value, high-traffic products that drive store revenue

Our approach for detecting promotions used price per unit fluctuations. If a product's sale price was significantly lower (≥30%) than its average price, it was flagged as promoted. To classify promotion trends, products were grouped based on the ratio of promoted transactions:

- Always Promoted (≥50% of transactions promoted)
- Moderately Promoted (5%–50%)
- Seldom Promoted (≤5%)



Promotion Analysis Summary:

3 rows | Length: 3, dtype: int64

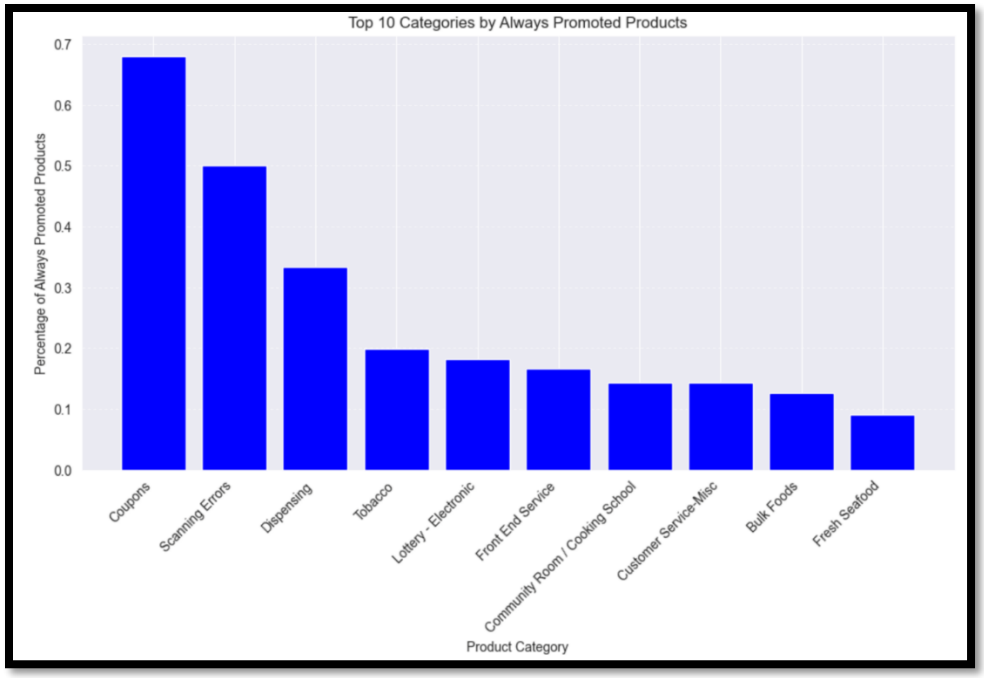
promotion_category	<unnamed>
Always Promoted	1525
Moderately Promoted	8861
Seldom Promoted	44966

Key Findings:

- A large majority of products (44,966) are seldom promoted
- Moderately promoted products (8,861) indicate occasional discounts but no consistent promotions
- Only 1,525 products are frequently promoted, suggesting a selective discounting strategy

Furthermore, we analyzed the top promoted products based on their promotion frequency, calculated by: promotion_analysis["promo_transactions"] / promotion_analysis["total_transactions"].

This visualization highlights the **top 10 categories with the highest percentage of always promoted products**.

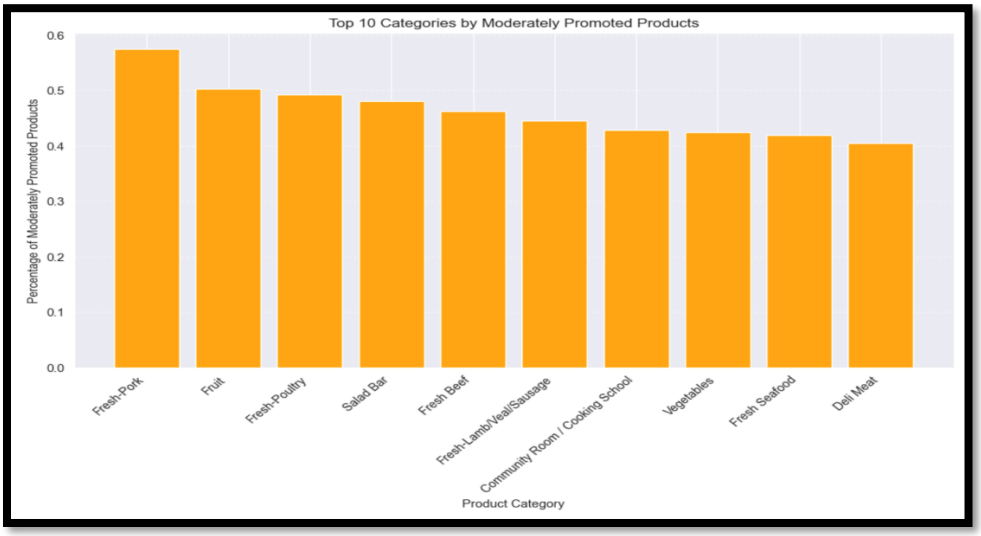


Key Findings:

- Coupons, scanning errors, and dispensing-related items have the highest percentage of always promoted products
- Tobacco, electronic lottery, and front-end service items are also consistently discounted
- Bulk foods and fresh seafood are among the few consumable categories in the top 10

These categories may be strategically discounted due to their nature (e.g., coupons, refunds, or high-margin products that sustain frequent promotions).

We also examined categories with the highest percentage of moderately promoted products.

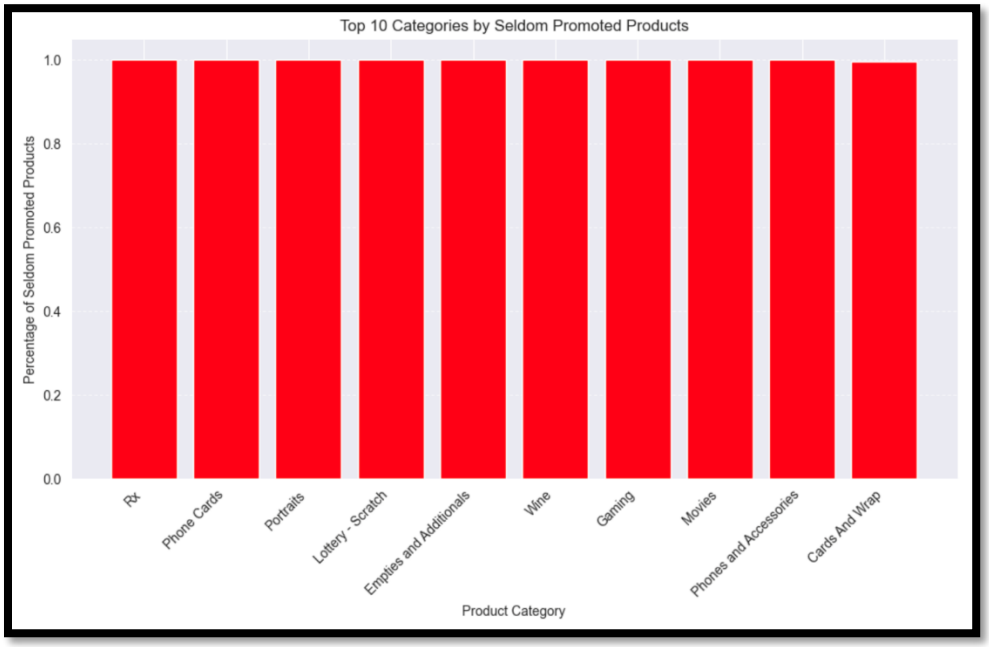


Key Findings:

- Fresh food items such as pork, poultry, beef, lamb, seafood, and vegetables are frequently moderately promoted
- Salad bars and deli meats also receive occasional promotions
- Cooking school and community room items are featured, suggesting discounts on services or classes

Retailers strategically promote fresh meat and produce to attract customers and drive foot traffic, as these items are essential and frequently purchased.

Additionally, we identified categories where products are seldom discounted.



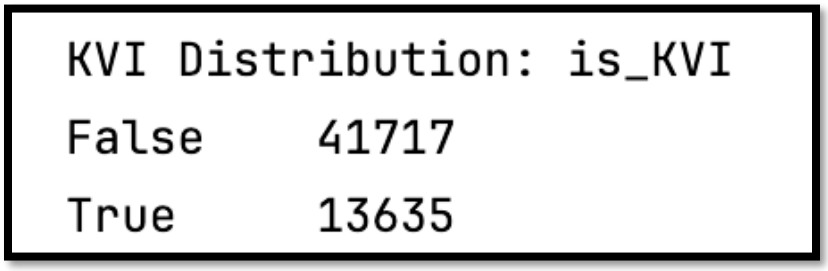
Key Findings:

- Pharmaceutical products (Rx), phone cards, and lottery tickets never go on promotion due to strict pricing regulations or fixed costs
- Luxury and entertainment items such as wine, gaming, and movies also rarely receive discounts, possibly due to manufacturer-controlled pricing or high demand stability
- Greeting cards and wrapping paper are other non-promoted items, suggesting low price elasticity in these segments

4.1.3 Key Value Items (KVI) and Key Value Categories (KVC) Analysis

Key Value Items (KVI) and Key Value Categories (KVC) are essential product groupings that drive revenue, transactions, and customer engagement. KVIs are high-demand products that influence store perception, while KVCs represent broader product groups containing high-value items.

Using a data-driven approach, we identified KVIs as products that ranked in the top 20% on either transactions or revenue.



Key Findings:

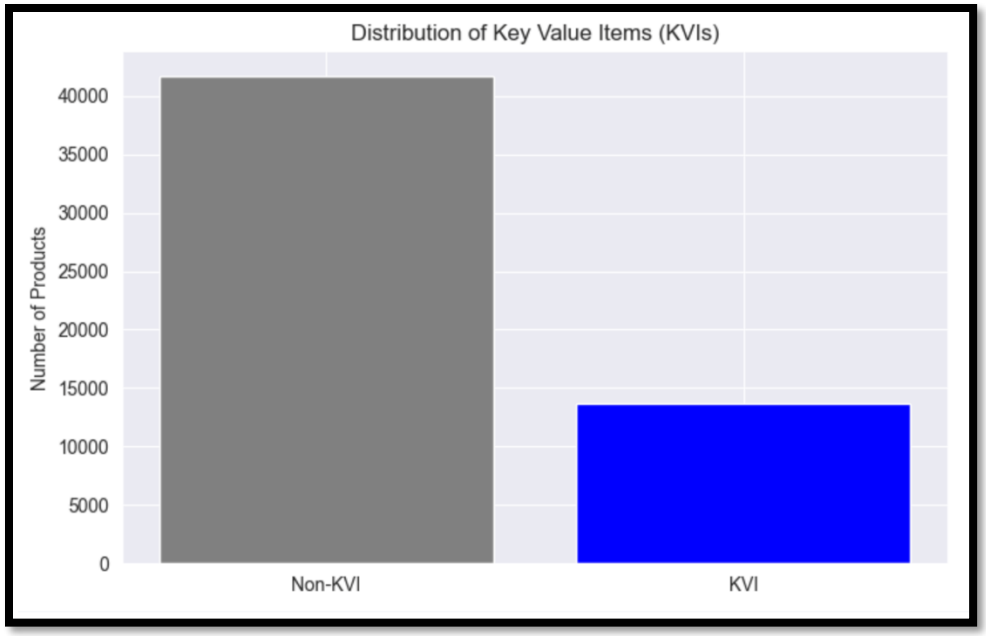
- Approximately 25% of all products are considered KVIs, representing the top-performing items
- KVIs generate 5x more revenue than Non-KVIs
- KVIs account for over 86% of all transactions, reinforcing their influence on store performance
- Non-KVIs have minimal impact on store traffic, emphasizing the need to prioritize KVIs in promotions and pricing strategies

Key Value Categories (KVCs) are product groups that contain a high proportion of Key Value Items (KVIs). Identifying KVCs helps businesses understand which product categories drive the most revenue and transactions, how to allocate promotions and pricing strategies efficiently, and which categories should be prioritized in inventory and stock management. This analysis classifies a category as a KVC if at least 30% of its products are KVIs.

is_KVI	total_units_sold	total_revenue	total_transactions
Non-KVI	202961	1091264.50	166815
KVI	1328103	5544493.16	1039248

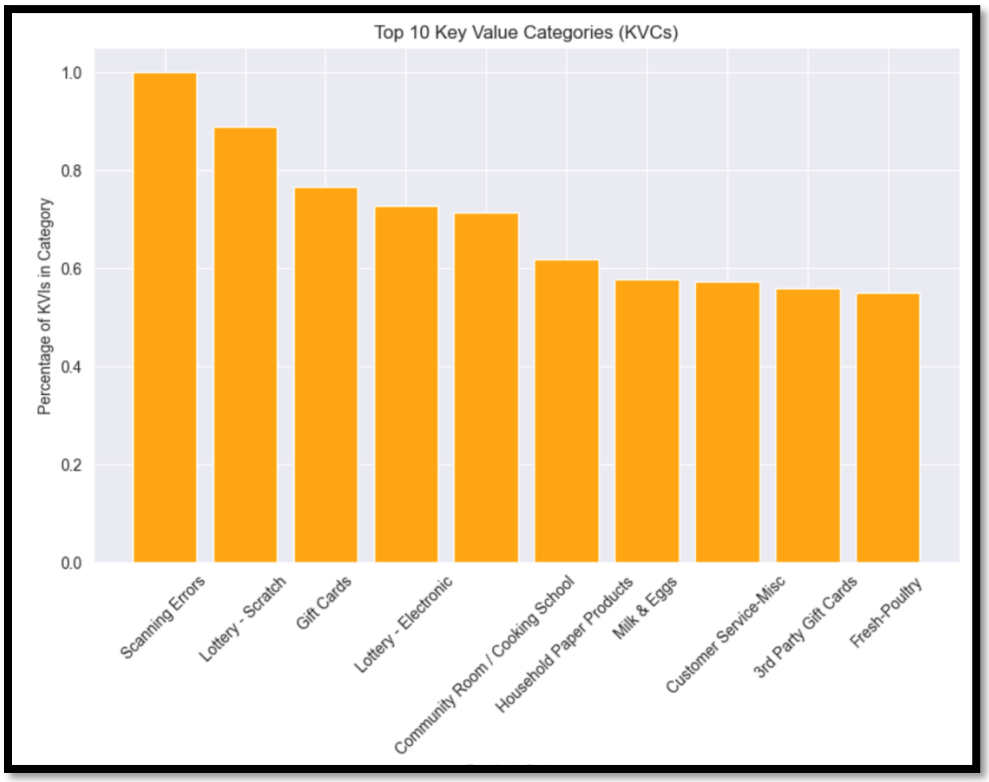
Key Findings:

- KVCs generate nearly 3 times the revenue of Non-KVCs, despite having fewer total products
- KVCs account for 75% of total transactions, confirming their importance in driving customer engagement
- Non-KVCs contribute significantly less to overall revenue, suggesting they are less critical in sales strategy



The majority of products (~41,000) are Non-KVIs, while only about 13,600 are KVIs. KVIs represent a smaller fraction of total products, but they significantly contribute to overall revenue and transactions.

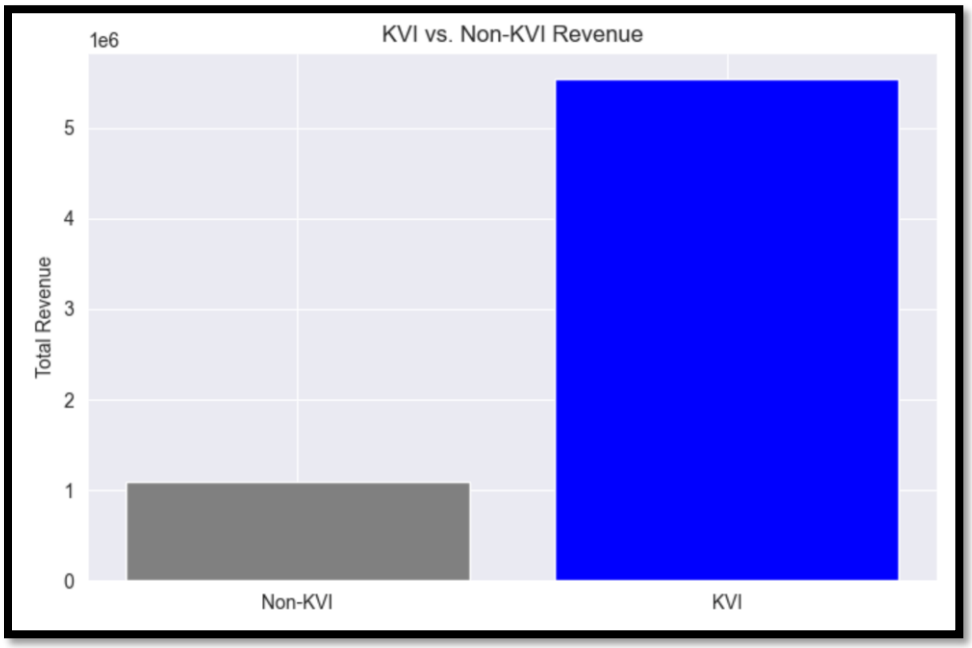
Bar chart showing categories with highest percentage of KVIs



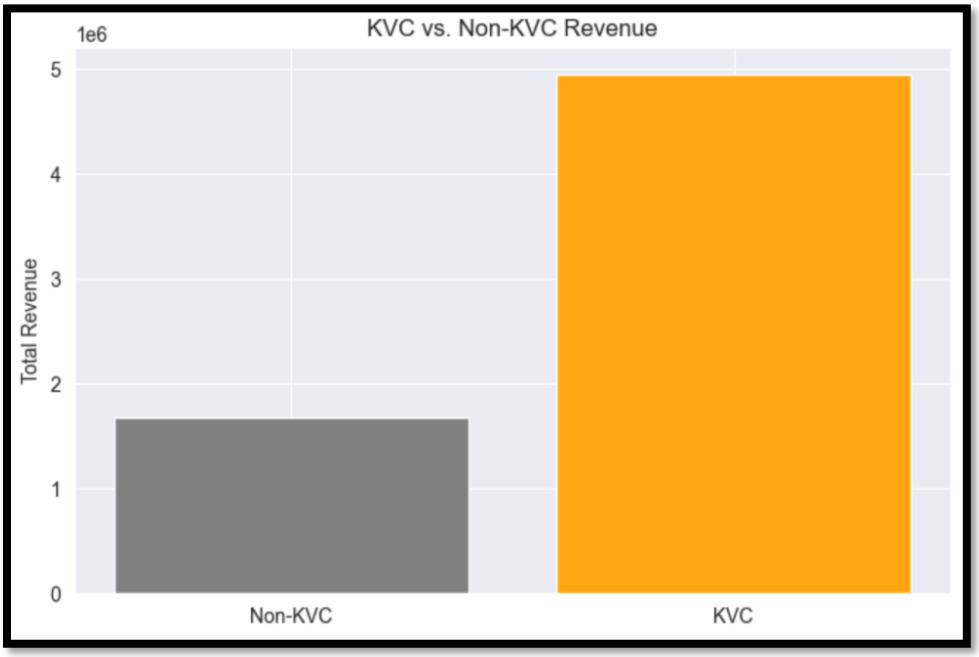
Key Findings:

- Scanning Errors and Lottery - Scratch categories contain nearly 100% KVIs, indicating these items are essential for customer purchases
- Gift Cards, Household Paper Products, and Milk & Eggs also have a high percentage of KVIs, suggesting these products are frequently purchased by consumers

Comparison chart showing revenue from KVIs vs Non-KVIs



Comparison chart showing revenue from KVCs vs Non-KVCs



Key Findings:

- KVIs generate significantly more revenue than Non-KVIs, despite being fewer in number
- Revenue from KVIs is nearly 5 times higher than Non-KVIs, confirming that KVIs are the backbone of the sales strategy
- KVCs dominate total revenue, reinforcing the importance of tracking and managing these categories
- Revenue from KVCs is significantly higher than non-KVCs, indicating that promotional and stocking strategies should focus on these high-performing categories

4.1.4 Conclusion on Product Analysis

Beyond traditional product categories and sub-categories, alternative groupings such as Key Value Items (KVIs), Key Value Categories (KVCs), and promotion-based classifications offer deeper insights into product performance. KVIs, which represent the top 20% of products by revenue or transactions, drive a disproportionately high share of sales and should be prioritized in inventory management and promotions. KVCs, which contain a high proportion of KVIs, demonstrate significantly higher revenue and transaction volumes than Non-KVCs, emphasizing their importance in strategic pricing and stocking decisions.

Additionally, promotion-based classifications reveal that most products are seldom promoted, while a small subset is consistently discounted to attract customers. These alternative product groupings provide valuable perspectives for optimizing pricing strategies, inventory allocation, and promotional efforts, ultimately enhancing revenue growth and customer retention.

4.2 Customer_Level Insights

4.2.1 Customer Purchase Behavior Analysis

Our analysis of customer purchase behavior revealed several significant patterns across various dimensions including spending, store visits, transaction frequency, and product variety. Using descriptive statistics, we gained a comprehensive understanding of overall customer behavior patterns.

Summary Statistics

Revenue Patterns:

- The minimum value of total_revenue is -469, which suggests the presence of refunds
- For better revenue tracking, we separated refunds to ensure that returned items do not distort customer value
- The mean total_revenue is 9.40, indicating that customers on average spend about $9.40
- The standard deviation is 28.65, which is quite high, meaning that spending varies

significantly among customers

Store Visit Behavior:

- The mean total_stores_visited is 1.09, with a median of 1, implying most customers visit only one store
- The maximum number of stores visited is 11, indicating some customers shop at multiple locations

Transaction Frequency:

- The mean total_transactions is 1.69, suggesting most customers make between 1 and 2 transactions
- The maximum number of transactions is 77, meaning there are a few highly active customers

Product Variety:

- The mean unique products purchased is 1.70, meaning most customers buy around 1 or 2 different products
- The maximum unique products purchased is 76, suggesting some customers purchase a very diverse range

4.2.2 High-Value Customer Analysis

To better understand our most valuable customers, we conducted detailed analyses of the top performers across multiple dimensions.

Top Customers by Revenue

We analyzed the top 1% of customers based on total revenue, which includes 6,956 customers.
Key observations:

- Top customer (cust_id: 60003028443430) spent $19,000, significantly higher than others
- This customer only visited one store, made one transaction, and bought one unique product
- This pattern might indicate a bulk purchase or a high-value item
- The lowest in the 99th percentile spent around $56.19
- There is a huge variance in spending behavior within the top 1%
- This suggests a mix of high-ticket buyers and frequent moderate spenders

Top Customers by Transaction Frequency

We analyzed the top 1% of customers based on transaction frequency, which includes 6,963 customers.
Key findings:

- The most frequent shopper (cust_id: 1147458804) made 77 transactions
- They visited 3 stores, purchased 73 unique products, and spent $356.78

- This suggests a diverse shopping pattern, possibly a loyal customer who frequently buys multiple products
- The second most frequent shopper (cust_id: 1135252713) made 76 transactions, purchasing 76 different products across 4 stores
- Their spending was $476.27, showing a high level of engagement
- The lowest transaction count in this group is 8 transactions, with spending as low as $34.57
- Some customers make many transactions with relatively low total spending while others make frequent transactions while purchasing a high variety of products

Top Customers by Store Visits

We analyzed the top 1% of customers based on the number of unique stores visited, which includes 55,219 customers.
Key insights:

- The customer who visited the most stores (cust_id: 1126005718) shopped at 11 different locations
- They made 22 transactions, purchased 22 unique products, and spent $119.32
- This suggests they are a highly mobile shopper, possibly preferring variety or convenience across locations
- The second most mobile shopper (cust_id: 1127306013) visited 10 stores, making 56 transactions and purchasing 55 unique products
- Their total spending was $277.01, showing frequent and diverse purchasing behavior
- The minimum store visits in the top 1% category is 2 stores, with spending as low as $4.58

This highlights a wide range of store visitation patterns, from multi-location shoppers to occasional store-hoppers. There is significant variation in transaction volume and spending patterns—some high store-visit customers make many small purchases across different locations, while others make larger purchases while frequently switching stores.

Top Customers by Product Variety

We analyzed the top 1% of customers based on the number of unique products purchased, which includes 7,106 customers.
Key observations:

- The customer with the highest product variety (cust_id: 1135252713) purchased 76 different products
- They made 76 transactions, visited 4 stores, and spent $476.27
- This suggests they are a diverse shopper, trying many different products
- The second highest variety shopper (cust_id: 1147458804) purchased 73 unique products
- They had 77 transactions across 3 stores, with a total spend of $356.78
- Their high transaction count and variety indicate they might be an explorer or bulk buyer
- The minimum product variety in the top 1% category is 8 unique products, with

spending as low as $32.27

There is a wide range of behavior, from moderate-variety shoppers to extreme variety-seeking customers.

4.2.3 Customer Segmentation Strategies

To better understand our customer base and develop targeted marketing strategies, we categorized customers into distinct segments based on several behavioral dimensions.

Spending-Based Segmentation

To categorize customers based on spending behavior, we defined three key segments:

- **Most Valuable Customers (MVCs)**: Top 1% of spenders
  - Customers in the 99th percentile and above based on total purchases
  - Total customers in this segment: 6,956
- **Regular Spenders**: Middle-tier spenders
  - Customers falling between the 25th and 99th percentiles of spending
  - Total customers in this segment: 514,419
- **Budget Buyers**: Lowest spenders
  - Customers in the bottom 25% of total purchases
  - Total customers in this segment: 173,837

Shopping Frequency Segmentation

To categorize customers based on shopping frequency, we defined the following segments:

- **One-Time Buyers**: Customers who made only one transaction
  - Total customers in this segment: 471,412
  - These shoppers may have made a single purchase and never returned
- **Occasional Shoppers**: Customers who made multiple transactions but are not frequent shoppers
  - Total customers in this segment: 216,837
  - These shoppers make purchases occasionally but not at a high frequency
- **Frequent Shoppers**: Top 1% of customers based on transaction count
  - Total customers in this segment: 6,963
  - These are the most engaged shoppers, making purchases frequently

Product Variety Based Segmentation

To categorize customers based on the diversity of products they purchase, we defined the following segments:

- **Single-Product Buyers**: Customers who buy only one unique product
  - Total customers in this segment: 468,557
  - These shoppers stick to a single product, possibly for necessity or loyalty

- **Moderate Variety Shoppers**: Customers who buy multiple products but are not in the top 1% for variety

    - Total customers in this segment: 219,549
    - These shoppers have a moderate level of variety in their purchases

- **Diverse Shoppers**: Top 1% of customers based on the number of unique products purchased
    - Total customers in this segment: 7,106
    - These are the most experimental and diverse shoppers, buying a wide range of products

Price Sensitivity Segmentation

To segment customers based on purchasing behavior, we defined the following categories:

- **Cherry-Pickers**: Customers who make small transactions, shop infrequently, but buy a wide variety of products
    - Total customers in this segment: 80
    - These shoppers likely seek deals and prefer variety over loyalty

- **Full-Price Buyers**: Customers who consistently spend more per transaction, shop frequently, and stick to specific products
    - Total customers in this segment: 154
    - These shoppers are less price-sensitive and tend to buy the same products regularly

- **Mixed Buyers**: Customers who fall in between the two extremes
    - Total customers in this segment: 694,978
    - This is the majority of shoppers, meaning most customers show no strong bias toward promotions or full-price loyalty

4.2.4 Category Preference Analysis

Analyzing customer preferences by product category revealed distinct patterns in shopping behavior:

Top Product Categories (Most Popular Among Customers):

- Vegetables: 78,557 customers – Many shoppers prioritize fresh produce
- Fruit: 56,452 customers – Strong demand for fresh, natural foods
- Natural Foods: 36,540 customers – Likely health-conscious shoppers
- Snacks: 33,496 customers – Impulse purchases, convenience-focused buyers
- In-Store Purchases: 25,521 customers – Likely checkout or front-end items

4.2.5 Key Customer Insights and Implications

Our comprehensive customer analysis revealed several critical insights with significant business implications:

1. **High Customer Concentration in Value**

- A small percentage of customers (1%) contributes disproportionately to revenue
- These high-value customers exhibit distinct shopping patterns worthy of specialized attention
- Targeted retention strategies for these customers could have outsized impact on overall revenue

2. **One-Time Buyers Represent a Majority**
   - Nearly 70% of customers make only one transaction
   - This represents a significant opportunity for conversion to repeat customers
   - Effective first-time buyer experiences and follow-up marketing could substantially improve retention

3. **Limited Store Exploration**
   - Most customers (>90%) shop at only one store
   - Cross-location promotions could encourage store exploration
   - The mobile segment (customers who visit multiple stores) shows higher engagement across multiple metrics

4. **Product Variety Correlates with Engagement**
   - Customers who purchase a wide variety of products tend to be more engaged shoppers
   - Encouraging product exploration may increase overall customer value
   - Personalized recommendations based on purchase history could expand customer product horizons

5. **Price Sensitivity Varies Significantly**
   - Clear segments exist from promotion-sensitive "cherry-pickers" to full-price loyal customers
   - Targeted promotion strategies can be developed for each segment
   - The majority of customers exhibit mixed buying behavior, requiring nuanced marketing approaches

These insights provide a foundation for developing targeted marketing strategies, personalized promotions, and enhanced customer experiences tailored to the specific needs and behaviors of different customer segments.

4.3 Store-Level Insights

4.3.1 Store Performance Analysis

The performance of ACSE stores was evaluated using key sales and customer metrics to identify the top-performing and underperforming locations. Using a comprehensive approach, we analyzed stores based on:

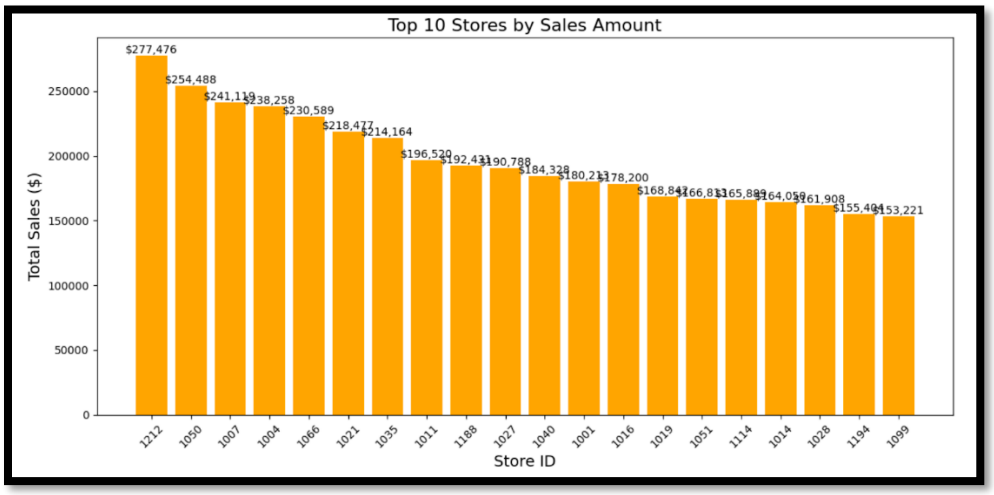
- Total Revenue: Sum of all sales (sales_amt) per store

- Total Transactions: Number of unique transactions (trans_id) per store
- Unique Customers: Number of distinct customers (cust_id) who made purchases in each store
- Total Sales Volume: Total number of product units sold (sales_qty)

This multidimensional approach provided a holistic view of store performance beyond simple revenue rankings.

Table showing sample store performance data

Index	*store_id*	*total_transactions*	*unique_customers*	*total_revenue*	*total_sales_volume*
0	1000	27599	20512	135047.39	33553
1	1001	28490	16860	180212.61	37040
2	1002	19264	13374	113931.41	25535
3	1003	41969	25477	238258.43	53400
4	1004	25035	15236	142370.34	32452



4.3.2 Store Segmentation Methodology

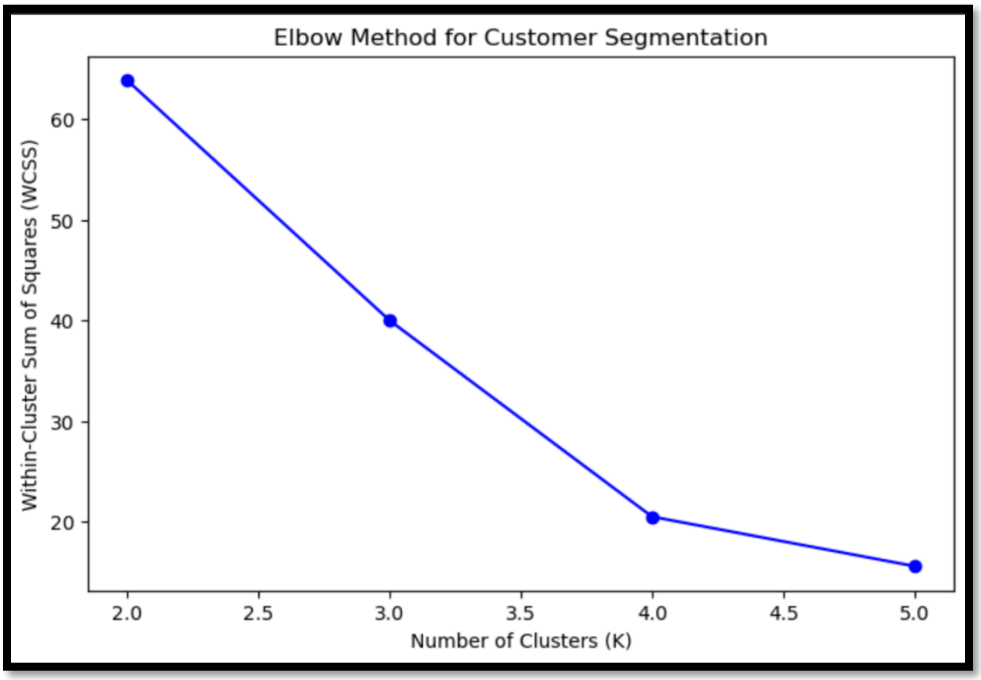
To better understand store performance patterns, we employed K-Means clustering using the key metrics: Total revenue, Total transactions, Unique customers, and Total sales volume. This unsupervised learning approach allowed us to identify natural groupings within our store network.

Since K-Means is sensitive to scale differences between variables, we first applied StandardScaler to normalize the data before clustering. This preprocessing step ensured that all metrics contributed equally to the clustering process, regardless of their original units or scales.

To determine the optimal number of clusters (K), we used the Elbow Method, which calculates the within-cluster sum of squares (WCSS) for different values of K and plots them

against the number of clusters. The point where adding more clusters no longer significantly reduces WCSS is considered the optimal K.

Elbow method plot showing WCSS for different K values



The results showed:

- K=3: Too broad, failing to differentiate underperforming stores effectively
- K=5: Introduced additional complexity but did not significantly improve segmentation
- K=4: Optimal choice, providing a well-balanced separation between different store types

Cluster visualization for K=3

Cluster visualization for K=5

Cluster visualization for K=4

The visual plot of the Elbow Method confirmed that K=4 was the ideal choice, as the decline in WCSS plateaued beyond this point.

4.3.3 Store Cluster Analysis

Using K=4, stores were segmented into four distinct groups with the following characteristics:

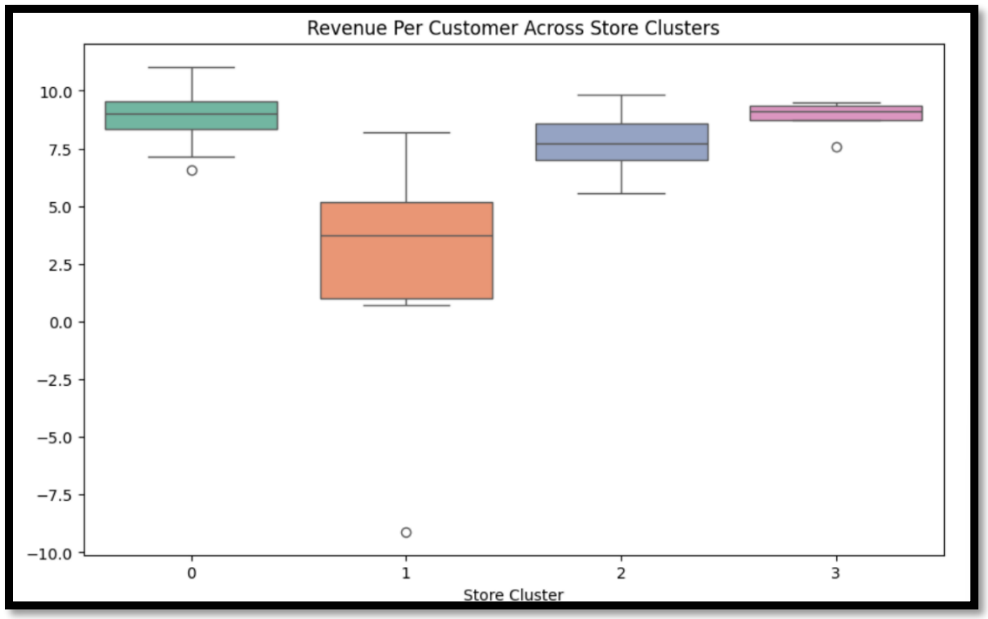
Table showing cluster statistics by revenue, transactions, customers, etc.

Cluster	*Avg Revenue*	*Avg Transaction*	*Avg Customers*	*Avg Sales Volume*	*Revenue per Customer*	*Store Count*
0	168427	29657	18852	38190	8.93	22
1	704	145	140	167	5.04	14
2	88346	16673	11156	21229	7.92	19
3	24386	44637	28193	57037	8.81	5

Table showing customer shopping behavior by store cluster

Cluster	*Avg_basket_size*	*Avg_revenue_per_transaction*	*Avg Customers*	*Avg_purchase_freq*	*store_numb*
0	1.044840	2.692255	2.69	1.004893	14
1	1.283824	5.603148	5.60	1.553197	25
2	1.281628	5.636071	5.64	1.655563	7
3	1.273566	5.210097	5.21	1.458354	14

Plot showing Revenue Per Customer across store clusters



The four clusters revealed distinct store profiles:

- **Cluster 0: Loyalty-driven stores (22 stores)**
  - Avg Revenue: $168,427
  - Avg Transactions: 29,657
  - Avg Customers: 18,852
  - Revenue per Customer: $8.93

    - Characterized by large basket sizes and high revenue per customer

- **Cluster 1: Underperforming stores (14 stores)**
    - Avg Revenue: $704
    - Avg Transactions: 145
    - Avg Customers: 140
    - Revenue per Customer: $5.04
    - Characterized by low revenue and low transaction volume

- **Cluster 2: High-traffic, high-revenue stores (19 stores)**
    - Avg Revenue: $88,346
    - Avg Transactions: 16,673
    - Avg Customers: 11,156
    - Revenue per Customer: $7.92
    - Characterized by moderate sales volume but strong revenue per customer

- **Cluster 3: Premium product stores (5 stores)**
    - Avg Revenue: $24,386
    - Avg Transactions: 44,637
    - Avg Customers: 28,193
    - Revenue per Customer: $8.81
    - Characterized by highest transaction volume and customer count

4.3.4 Strategic Recommendations by Store Cluster

Based on store rankings and segmentation, we developed cluster-specific strategic recommendations:

For Cluster 0 (Loyalty-Driven Stores)
- Maintain high-value customer engagement through loyalty programs
- Focus on exclusive promotions to retain frequent high-spending customers

For Cluster 1 (Underperforming Stores)
- Investigate factors contributing to low performance (e.g., product mix, location, pricing)
- Implement targeted discount campaigns to increase store visits

For Cluster 2 (High-Traffic, High-Revenue Stores)
- Ensure proper inventory restocking and supply chain optimization to prevent stockouts

Introduce cross-selling and bundling strategies to increase transaction values

For Cluster 3 (Premium Product Stores)
- Optimize pricing and promotions for premium product lines
- Enhance in-store customer experience to increase premium product conversion rates

These targeted strategies acknowledge the unique characteristics of each store cluster and provide specific approaches to optimize performance across the entire store network.

5. Recommender System & Market Basket Analysis

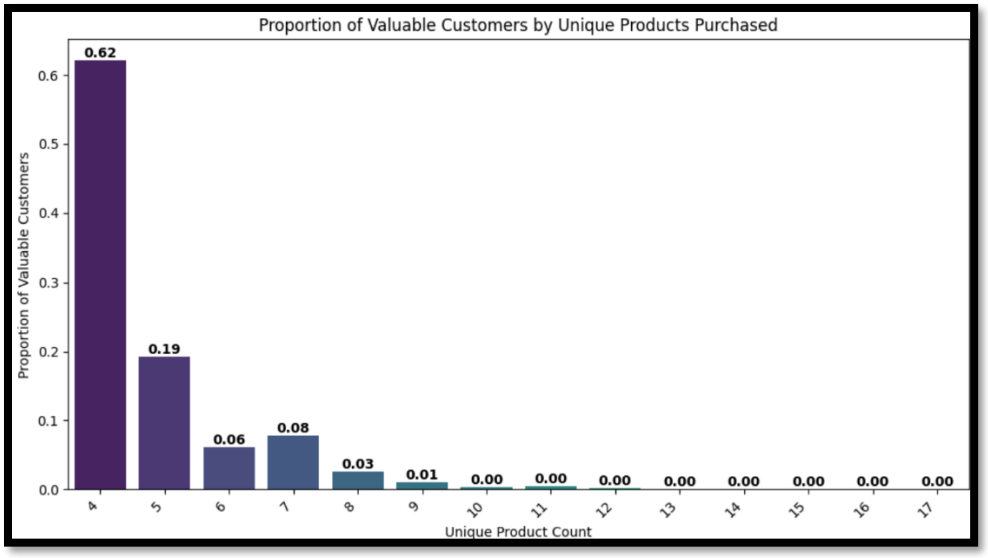
5.1 Recommender System Development

Implementing Collaborative Filtering (User-Based & Item-Based)

Most Valuable Customers: Top 10% of spenders, likely to purchase premium products.

- These customers spend significantly more than others.
- They purchase a wide variety of products, showing diverse shopping habits.
- Discount-driven shoppers (promo_ratio < 0.3) are excluded to focus on those who buy at regular prices.
- Customers are sorted by total spending to rank the most valuable ones.

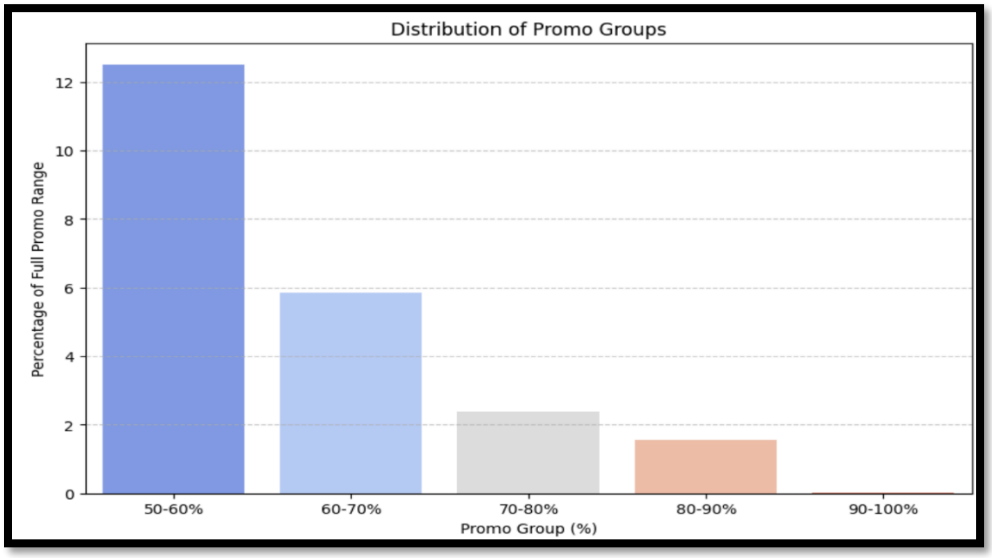
Chart showing Most Valuable Customers' spending patterns



Promo Sensitivity-Based group

- Customers were divided into five promo sensitivity groups, ranging from low promo reliance (0-50%) to extreme promo reliance (90-100%).
- A significant portion of customers fall into the 60-70% promo range, indicating moderate price sensitivity.
- The 90-100% promo group represents "cherry-pickers", who primarily buy on promotions.

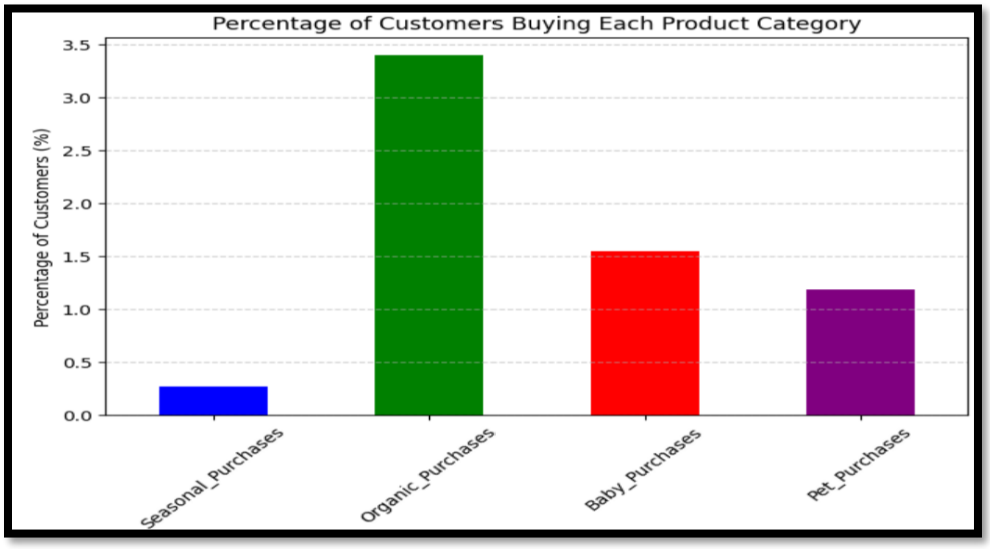
Graph showing customer distribution across promo sensitivity groups



Content-based recommendations using product descriptions. We choose four issue of product descriptions:

- Organic products have the highest customer engagement (~3.5%), indicating a strong demand for health-conscious and eco-friendly products.
- Baby and pet products also show significant buyer interest (~1.5% and ~1.2%, respectively), suggesting opportunities for cross-promotions and subscription models.
- Seasonal products have the lowest engagement (~0.3%), implying that purchases in this category are event-driven rather than consistent throughout the year.

Bar chart showing customer engagement with different product categories

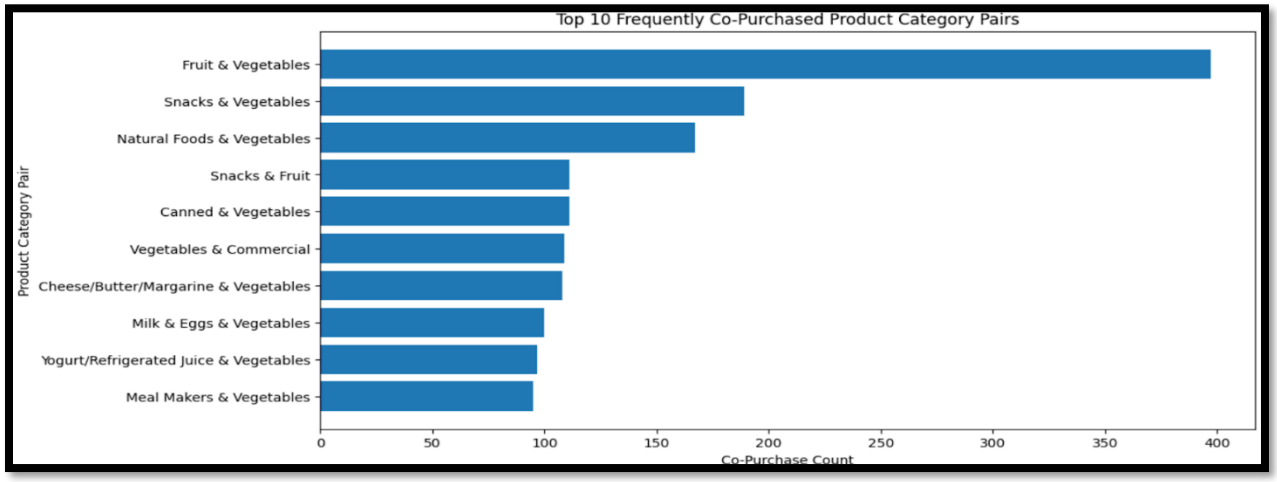


5.2 Market Basket Analysis

Identifying frequently purchased product combinations.

Top 10 Frequently Co-Purchased Product Category Pairs:

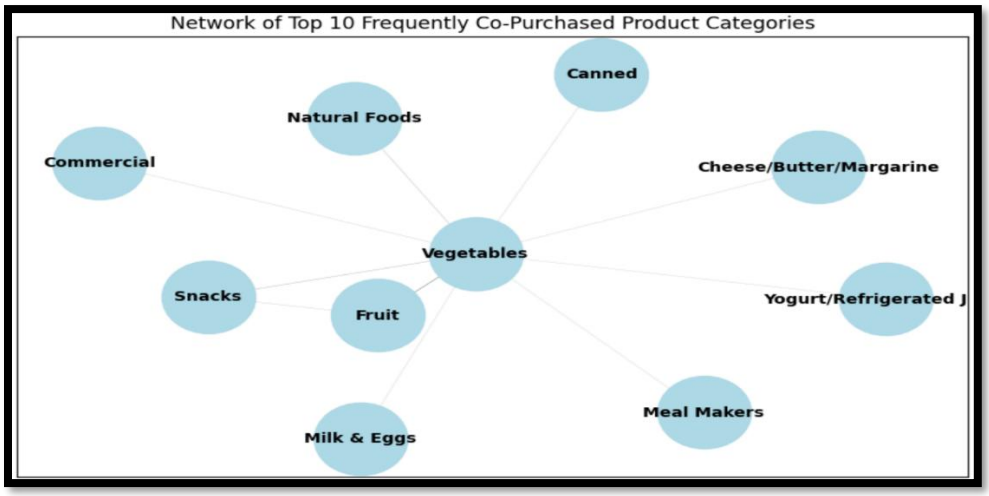
- This visualization highlights the most common product category pairs purchased together.
- The most frequent co-purchased pair is "Fruit & Vegetables", followed by "Snacks & Vegetables" and "Natural Foods & Vegetables".
- This suggests that fresh produce is often bought with complementary items like snacks and canned goods.



Finding cross-sell opportunities through association rule mining.

Network Graph of Co-Purchased Categories:

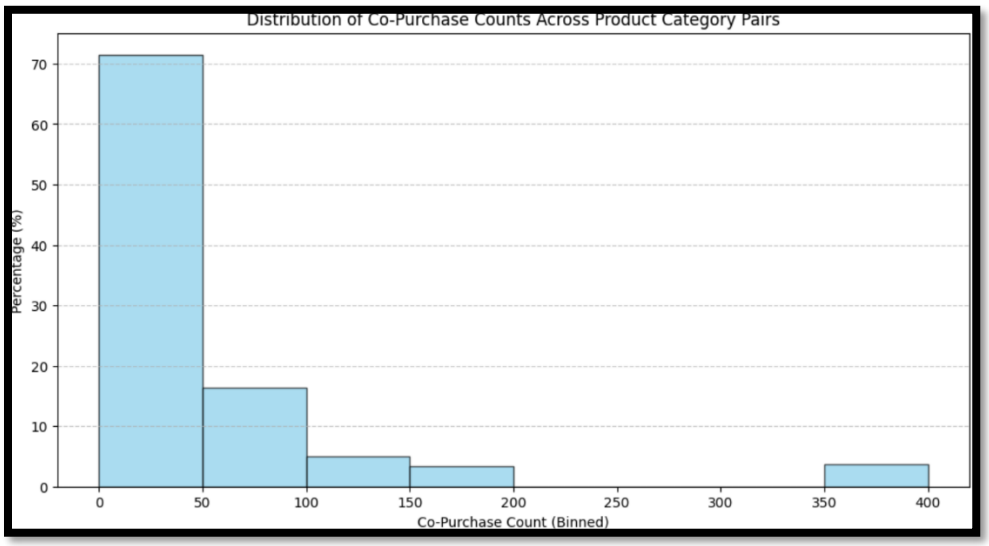
- This visualization shows a network of top co-purchased product categories, with "Vegetables" as the central node.
- Other frequent connections include Canned Foods, Dairy (Milk & Eggs, Cheese), and Beverages (Juice, Yogurt).
- This confirms that staple grocery items are often purchased together.



Distribution of Co-Purchase Counts:

- The x-axis represents the number of times a category pair was co-purchased, grouped into bins of 50.
- The y-axis represents the percentage of total category pairs falling within each bin.
- Over 70% of category pairs have low co-purchase counts, meaning most product pairings are unique or infrequent.
- A long-tail effect is observed, where a few category pairs have very high co-purchase frequencies.

Histogram showing distribution of co-purchase counts



Key Insights

- Most category pairs are bought together only a few times, indicating a diverse shopping pattern rather than strong repetitive purchases.
- A few dominant product pairs emerge, with "Vegetables" acting as a common connector to many categories.
- The long-tail distribution suggests rare product pairings exist, but a few key bundles drive most sales.

5.3 Insights & Recommendations

Product recommendations for enhancing customer experience.

- Most valuable customers exhibit high spending and diverse product purchases, making them ideal targets for personalized recommendations.
- Collaborative filtering is effective for engaged users but struggles with new users (cold start problem).

Promotions & Bundling:

- Identify high-frequency co-purchased pairs and offer bundled discounts.
- Example: Promote "Milk & Cereal" or "Vegetables & Natural Foods" bundles to encourage cross-selling.

Store Layout Optimization:

- Arrange frequently co-purchased items near each other to increase impulse buying.
- Example: Placing "Chips & Soda" together can lead to more combined purchases.

Targeted Marketing Campaigns:

- Use association rule insights for personalized recommendations.
- Example: If a customer buys "Pasta," suggest they also buy "Pasta Sauce" in online and in-store promotions.

Low-Frequency Items Strategy:

- Analyze products with low co-purchase counts and determine whether they need better placement, advertising, or pricing adjustments.

Market basket analysis provides valuable insights into customer buying behaviors, helping businesses optimize inventory, improve marketing strategies, and increase revenue through cross-sell opportunities.

6 Business Implications & Recommendations

6.1 Strategies for Optimizing Product Assortment and Pricing

Based on our comprehensive data analysis, we recommend the following strategies to optimize ACSE Supermarket's product offerings and pricing approach:

6.1.1 Key Value Item (KVI) Management

Our analysis revealed that approximately 25% of products (KVIs) generate 5x more revenue than non-KVIs and account for over 86% of all transactions. To capitalize on this finding, we recommend:

- **Prioritize inventory management for KVIs**: Ensure consistent availability of top-performing products, particularly those identified in our analysis like "Pharmacy Rx," bananas, and fresh produce items. Implement automated reordering systems that maintain higher safety stock levels for these critical items.

- **Strategic shelf placement for KVIs**: Position Key Value Items in high-visibility areas throughout the store, with particular attention to categories with high percentages of KVIs such as Gift Cards, Household Paper Products, and Milk & Eggs.

- **Premium positioning for non-discounted categories**: For categories we identified as rarely discounted (pharmaceuticals, lottery tickets, luxury items), focus on quality messaging and premium merchandising rather than price promotions.

6.1.2 Category-Specific Pricing Strategies

Based on our promotion analysis findings, we recommend tailored approaches for different product categories:

- **Fresh food rotation strategy**: Implement a strategic rotation of moderate promotions across fresh food items (meat, produce, seafood) that our analysis showed respond well to promotions and drive store traffic.

- **Bundle pricing for complementary items**: Create promotional bundles based on our market basket analysis, particularly focusing on the high-frequency co-purchase pairs like "Fruit & Vegetables," "Snacks & Vegetables," and "Natural Foods & Vegetables."

- **Selective premium product promotions**: For categories in premium product stores (Cluster 3), develop limited-time premium product promotions that maintain margin while driving trial.

6.2 Recommendations for Targeted Marketing and Personalized Customer Experiences

6.2.1 Customer Segment-Specific Marketing

Our customer segmentation analysis revealed distinct shopping patterns that can inform targeted marketing strategies:

- **MVCs retention program**: Develop an exclusive benefits program for the Most Valuable Customers (top 1% of spenders) who collectively drive significant revenue. Based on their identified purchase patterns, create personalized rewards that encourage continued high-value shopping behavior.

- **One-time buyer conversion strategy**: With 471,412 customers (68%) making only a single transaction, develop targeted re-engagement campaigns focusing on their initial purchase categories to drive a second visit.

- **Mobile shopper cross-store incentives**: For the 55,219 customers identified as shopping at multiple locations, create incentives that reward exploration of additional stores in the network.

6.2.2 Personalized Product Recommendations

Based on our market basket analysis and customer segmentation, we recommend:

- **Collaborative filtering for diverse shoppers**: Implement collaborative filtering recommendations for customers who purchase a wide variety of products, particularly focusing on the 7,106 customers identified as "Diverse Shoppers" who purchase in the top 1% of product variety.

- **Content-based recommendations for category specialists**: For customers who focus their purchases in specific categories, provide content-based recommendations that expand their selection within those categories, gradually introducing complementary categories.
- **Address cold start problem for new customers**: Develop a hybrid recommendation approach that leverages popularity data for initial interactions while quickly transitioning to more personalized recommendations as customer data accumulates.

6.2.3 Store Cluster-Specific Experience Enhancements

Our cluster analysis identified four distinct store types, each requiring tailored approaches:

- **Loyalty-Driven Stores (Cluster 0)**: For these 22 stores with large basket sizes and high revenue per customer, focus on enhancing the loyalty program experience with exclusive events and personalized services.

- **Underperforming Stores (Cluster 1)**: For the 14 stores with low transaction volume, conduct in-depth competitive analysis and develop location-specific revival strategies, potentially including revised product mix and targeted local marketing.

- **High-Traffic Stores (Cluster 2)**: For the 5 high-traffic locations, optimize operational efficiency and implement cross-selling initiatives to capitalize on the high customer flow.

- **Premium Product Stores (Cluster 3)**: For the 19 stores with strong revenue per customer, enhance the premium shopping experience through specialized staff training and curated product displays.

6.3 Potential Areas for Further Analysis and Experimentation

6.3.1 Advanced Analytics Opportunities

To further enhance ACSE's data-driven decision making, we recommend exploring:

- **Predictive churn modeling**: Develop algorithms to identify at-risk customers before they stop shopping with ACSE, particularly focusing on patterns that precede customer attrition.

- **Seasonal product trend analysis**: Conduct time-series analysis to identify and predict seasonal variations in product demand, especially for the seasonal products that showed lower but concentrated engagement patterns.

- **Customer lifetime value prediction**: Create models that estimate the long-term value of different customer segments to inform acquisition and retention investment decisions.

- **Market basket sequence analysis**: Expand our current market basket findings to include the sequence of purchases over time, identifying which products typically serve as entry points to different categories.

6.3.2 Recommended A/B Testing

To validate the effectiveness of our recommendations, we suggest conducting A/B tests for:

- **Store layout optimization**: Test alternative product placements based on our co-purchase findings, comparing traditional category-based arrangements against complementary product groupings.

- **Promotion format effectiveness**: Experiment with different promotion structures (percentage discounts vs. bundle pricing vs. loyalty points) across different customer segments to identify optimal approaches.

- **New customer welcome journeys**: Test various onboarding experiences for first-time shoppers to determine which approaches most effectively convert them to repeat customers.

- **Cherry-picker value proposition testing**: Experiment with different strategies to shift the shopping behavior of extreme promotion-reliant customers toward more profitable patterns.

By implementing these data-driven strategies and continuously testing new approaches, ACSE Supermarket can leverage its rich customer and product insights to enhance the shopping experience, optimize operations, and ultimately drive sustainable revenue growth and profitability.