ACSE Supermarket Recommender System Report

AI & ML at Scale Team 1

Lanston Chen, Songbo Hu, Kedi Lin, Jie Mei, Jayson Xu, Wenxi Xu

Emory Goizueta Business School

Executive Summary

This report aims to deliver a sophisticated analysis encompassing data understanding, cleansing, and exploratory examination to uncover insights pivotal for guiding strategic initiatives at ACSE Supermarket, thereby elevating its operational efficacy and market standing. Our analytical endeavors will dissect provided datasets to unravel the intricacies of customer demographics, product performance, and store efficiencies. Objectives focus on identifying key customers, products, and stores, uncovering distinct segments, and operational insights. The outcomes of this analysis are intended to fortify ACSE Supermarket's decision-making in supply chain optimizations, operational enhancements, pricing models, marketing strategies, and promotional frameworks.

Business Understanding

Problem Definition:

ACSE Supermarket operates in a highly competitive retail environment, offering a vast array of products to a diverse customer base across over 40 stores in North America. As part of its strategic initiative to enhance business performance and customer satisfaction, ACSE aims to leverage advanced analytics to inform its decision-making processes. The introduction of a recommender system is envisioned as a cornerstone of this strategy, aimed at optimizing various aspects of the business, including supply chain logistics, store operations, inventory management, and customer engagement.

The challenge lies in effectively analyzing ACSE's extensive dataset, which encompasses transaction histories and product information for over 100,000 items across more than 100 categories. This analysis must achieve several critical objectives:

- **Customer Insight**: Identify high-value customers by analyzing transaction data and customer interactions. Understand their purchasing patterns, preferences, and behaviors to tailor marketing strategies and product offerings.
- **Product Performance**: Evaluate the sales performance of products and categories to determine which items contribute most significantly to revenue and profitability. This involves understanding which products are frequently purchased together and identifying trends and seasonal variations in product popularity.
- **Store Optimization**: Rank stores based on various performance metrics, including sales volume, revenue, profitability, and customer traffic. Identify

factors contributing to the success of top-performing stores to replicate these strategies across the network.

- **Market Trends**: Uncover emerging trends in customer preferences and market demand to anticipate future product needs. This includes identifying underperforming areas and potential growth opportunities.
- **Data Integrity**: Addressing data integrity by excluding specific categories such as plastic bags, pharmaceuticals, tobacco, and alcohol due to legal and model requirements and correcting null values, anomalies, abnormal distributions, and input errors.

The development of a recommender system is expected to address these challenges by providing actionable insights that can inform strategic decisions across ACSE's operations. This system must be capable of processing and analyzing complex datasets to recommend products to customers based on their purchase history and preferences, thereby increasing sales and customer loyalty. Additionally, it should offer valuable guidance on inventory management, store layout optimization, and marketing strategies, contributing to an enhanced shopping experience and operational efficiency.

Expected Outcomes:

- **Increased Sales and Revenue**: By recommending products that align with individual customer preferences and historical purchasing patterns, ACSE expects to see an uplift in sales. Specifically, targeted product recommendations are anticipated to boost cross-selling and upselling opportunities, thereby increasing average transaction values.
- **Inventory Optimization**: The recommender system will enable more accurate demand forecasting for each store, leading to optimized stock levels. This means reducing overstock and understock situations, which in turn reduces inventory holding costs and minimizes lost sales due to stockouts. Specific outcome metrics include improved turnover rates and reduced markdowns on overstocked items.
- **Enhanced Customer Experience**: Personalized recommendations are expected to enhance the shopping experience for ACSE's customers, making it easier for them to find products they love and perhaps discover new items. This personalized approach should lead to higher customer satisfaction and loyalty, as measured by repeat visitation rates and an increase in the Net Promoter Score (NPS). NPS will be calculated in the future model.

- **Improved Marketing ROI**: With the help of the recommender system, marketing campaigns can be more accurately targeted, leading to higher conversion rates and more efficient use of marketing budgets. Specific outcomes include higher engagement rates on promotional emails and increased effectiveness of online and in-store promotions.
- **Strategic Supplier Negotiations**: By understanding product performance and customer preferences in greater detail, ACSE can enter supplier negotiations with better insights. This could result in more favorable terms, exclusivity deals, or collaboration on product development, directly impacting the cost of goods sold (COGS) and product availability.

Data Understanding

Data Overview:

The dataset from ACSE Supermarket is structured into two key tables:

Transactions Table

This table records detailed transaction histories from 2017 to 2020, spanning over 1.2 billion rows, which reflect the supermarket's extensive customer interactions. Each transaction entry includes a customer ID, store ID, product ID, transaction ID, and date, alongside quantitative details like sales quantity, weight (for products sold by weight), and the sales amount before discounts. The transaction table's significance lies in its capacity to:

- Trace customer purchase paths, revealing patterns in buying behavior, preferences, and loyalty trends.
- Measure store-specific performance metrics, including foot traffic, sales volume, and revenue generation, to identify high-performing locations and areas needing improvement.
- Analyze product demand and sales trends, aiding in inventory management, and identifying key revenue drivers within ACSE's product range.

Products Table

Containing information on over 100,000 products, the products table delves into the specifics of each item sold, including product ID, description, section, category,

subcategory, type, manufacturer or brand code, unit quantity count, and the unit of measure in 150,000 entries. This rich product metadata allows for:

- Detailed performance analysis at the product level, facilitating strategic decisions regarding product assortment, pricing, and promotions.
- Insight into product categorization, which supports marketing strategies and the optimization of store layout according to product sections and categories.
- Brand performance assessment, guiding supplier negotiations and brand partnerships.
- Initial Observations: Share initial insights or interesting findings from a preliminary examination of the data. Discuss the quality of the data, including any potential issues like missing values or inconsistencies.
- Data Exploration Plan: Outline the steps for thorough data exploration, including statistical analyses, data visualization, and any specific techniques to uncover patterns or anomalies in the data.

Data Preparation

Data Cleaning:

Customer Data Preparation

- **Date Filtering**: Transactions recorded after 2020-03-01 are treated as COVID data. We excluded it to maintain consistency in the analysis period and focus on pre-COVID data unaffected by external global events.
- **Activity Threshold**: To concentrate on engaged customers, those who visited the store less than 5 times in total or spent less than $100 overall are filtered out, highlighting significant customer interactions.
- **Frequency Limit**: Transactions indicating more than 10 visits by a single customer in a day are considered outliers and excluded to ensure a realistic representation of shopping behavior.

Product Data Preparation

- **Plastic Bags**: Given its status as the most sold item, transactions involving product ID 20189092 (plastic bags) are excluded from the recommender system analysis. This decision is based on the rationale that plastic bag sales are a

byproduct of other purchases, making them impractical for recommendations. However, for profitability analysis, plastic bags are retained due to their significant profit contribution and pricing flexibility.

- **Gift Cards**: Excluded due to the complexity of profit calculation. Gift cards carry minimal cost but can significantly influence profitability indirectly for store-issued cards and directly through fixed commissions for third-party cards. The phenomenon of unused card balances eventually contributing to profits further complicates their profitability analysis.
- **Other Categories**: Transactions categorized under front-end service, scanning errors, customer service-misc, empties, and additionals are excluded. These categories are removed due to their non-standard nature, which could distort the analysis of typical product sales and customer behavior.

Invalid Transactions

- Transactions not meeting logical criteria for sales data are filtered out to maintain data quality. **This includes:**
    - Transactions with positive sales amounts and negative quantities/ weights that are not logically interpretable.
    - For both positive sales amounts (normal sales) and negative sales amounts (return), sales quantities and sales weights are generally treated as mutually exclusive metrics. However, for instances where the quantity equals 1, we permit the coexistence of both metrics in the data. All other scenarios where both metrics are present but do not meet this specific condition are excluded.
    - Also, exclude transactions with abnormal amounts, quantities, and weights that are inconsistent with their designated labels (Coupon, Returns).

Data Transformation

The original dataset contains over 9 million users' data from the store. After cleaning the data, the total customer number is **3,329,518**. We've chosen to sample 10% of the customer data from the processed set of **826,992,297** transactions to facilitate text mining and model building efficiently. This customer-centric sampling strategy is adopted instead of transaction-based sampling to preserve each customer's history, enabling a deeper understanding of individual purchasing patterns. Due to technical and scale constraints on the Google Cloud Platform, merging the profit table with the original transactions for profit calculation is not feasible. This

approach balances computational efficiency with the goal of maintaining comprehensive insight into customer behavior within the constraints of our technical infrastructure.

More Discussions

In our discussions on data trimming strategies, particularly focusing on customer purchase frequency, we observed notable patterns. The highest frequencies of purchases in recent years within the dataset were 90265, 71181, and 63729, with data points above 20,000 being sparse (Appendix I). Based on this, we considered three distinct approaches, each with significant implications:

- **Traditional Trimming of the Top and Bottom 5%**: This method would yield a more normalized and consistent data structure but at the cost of significantly reducing the estimated profit margins for the store and losing a portion of the data. Besides removing outliers, this approach also eliminates a large number of continuous data points.
- **Exclusion of Data Above 20,000 Based on Observed Patterns**: This approach aligns more closely with the actual distribution of the data, with the criteria based on observations rather than following a fixed pattern. However, the selection of the cutoff point could significantly impact the final model's results.
- **No Action on the Upper Tail**: By not trimming the upper tail, we retain maximum information, allowing for an in-depth exploration of the most valuable customers. Since the dataset does not contain many outliers, leaving the upper tail untouched is unlikely to affect the results adversely.

For the lower tail, we chose to eliminate individuals with fewer than five total transactions and who spent less than $100, defining them as low-value customers. This group introduces randomness, justifying their exclusion. However, this decision brings additional considerations: due to the short timeline, newer customers haven't had the time/opportunity to accumulate sufficient visits or spending, potentially misclassifying their value compared to more established customers. This filter indiscriminately removes both new and long-term customers without distinguishing between them. It also does not consider the difference between loyal and casual customers. Our profit distribution analysis reveals that loyal customers significantly contribute to the supermarket's profits (Appendix II), yet our filters do not differentiate treatments between these groups, which may affect the accuracy of subsequent models and calculations.

Query Results

Due to the prohibitive size of the datasets, preventing integration with local tables through BigQuery, we limited our analysis to a 0.1% data sample. The justification for this sampling approach will be detailed subsequently. We compute the product's profit by multiplying its sales volume with the respective category's profit margin, utilizing retrieval and analysis of publicly accessible data across more than 100 product categories in the North American market.

Identifying Top-Tier Customers: A Revenue and Profit Perspective

	cust_id	total_revenue
0	1112173002	271612.77
1	1136587321	152553.31
2	1142991546	140632.96
3	1127629119	136476.82
4	1147882948	124585.03
5	60002346311410	121791.07
6	1130336013	117890.69
7	1136729077	106310.21
8	60003120066580	104923.37
9	1128632096	104759.70

Figure 1.1 Customer with the most revenue

	cust_id	transaction_count
0	60003120066580	19660
1	1130336013	19393
2	1128632096	17802
3	1124821673	16739
4	1126222672	15294
5	1129815034	14915
6	1130375348	14735
7	1151662833	14595
8	1128123822	14560
9	1123922650	14149

Figure 1.2 Customer with the most transaction count

	cust_id	visit_count
0	1125563829	911
1	1132869627	892
2	1131263734	877
3	1123859009	877
4	1125142670	876
5	1127503633	869
6	1125069207	867
7	1124797671	864
8	1125160896	857
9	1125617054	849

Figure 1.3 Customer with the most visit

cust_id	profit
1008740211	12178.41980
1124764327	10232.90645
1125265308	8371.51935
1128742376	8107.16355
1142684801	7456.56025
1143167569	7197.90830
1126371663	6973.89475
1127904558	6962.21885
1129917824	6877.76415
1136083110	6600.96960

Figure 1.4 Customer with the most profit

	cust_id	product_count
0	1130336013	6825
1	60003120066580	6814
2	1128632096	5798
3	1126222672	5487
4	1123922650	5440
5	1124985425	5172
6	1151662833	5118
7	60003133573240	5116
8	33212674139	5096
9	1128123822	5087

Figure 1.5 Customer with the most product count

Best-Selling Products and Groups

	prod_id	total_units_sold
0	20175355001	13720283
1	21097012001	6818371
2	20047851	6355216
3	20123850	6085925
4	20028593001	6055142
5	20040489001	5468571
6	20070132001	5240164
7	20076950	4774431
8	20668578	4353108
9	20812144001	4152426

Figure 2.1 Products with the most units sold

	prod_id	total_weight_sold
0	20175355001	13219141.36
1	20139509001	2478044.90
2	20425775001	2045943.10
3	20159199001	1919931.28
4	20426078001	1795505.24
5	20127708001	1615074.80
6	20026703001	1532684.46
7	20083526001	1462454.36
8	20426141001	1369735.68
9	20159690001	1104106.60

Figure 2.2 Products with the most weight sold

	prod_id	total_prod_trans
0	20175355001	13720294
1	20668578	4348227
2	20070132001	4242970
3	20812144001	3574510
4	20145621001	3102016
5	20007535001	3042138
6	21097012001	3035599
7	20527440	2934861
8	20671789	2920847
9	20028593001	2914054

Figure 2.3 Products with the most transactions

	prod_category	total_units_sold
0	Vegetables	122648828
1	Fruit	96113165
2	Snacks	56152072
3	Natural Foods	55434773
4	In-Store	45439939
5	Cold Beverages	42424542
6	Canned	40961901
7	Milk & Eggs	34638850
8	Meal Makers	33465730
9	Commercial	32018523

Figure 2.4 Product categories with the most units sold

	prod_category	total_weight_sold
0	Fruit	32499912.66
1	Vegetables	20022459.67
2	Fresh-Poultry	8617427.59
3	Fresh Beef	6392292.11
4	Fresh-Pork	2215236.32
5	Salad Bar	1535923.81
6	Deli Meat	1506839.74
7	Fresh Seafood	1299724.15
8	HMR	932988.05
9	Fresh-Lamb/Veal/Sausage	712270.27

Figure 2.5 Product categories with the most weight sold

	prod_category	total_prod_cate_trans
0	Vegetables	112446078
1	Fruit	76951653
2	Snacks	46568633
3	Natural Foods	44251003
4	In-Store	32576304
5	Cold Beverages	30781675
6	Milk & Eggs	30448381
7	Commercial	27714876
8	Meal Makers	25789376
9	Cheese/Butter/Margarine	23847899

Figure 2.6 Product categories with the most transactions

Figures 1.1 to 1.4 provide a clear visualization of customers ranking highest in specific aspects, offering us an intuitive understanding of which customers or stores contribute the most under different filtering criteria. This helps us grasp the outliers and the scale of the data. Figures 2.1 to 2.12 offer insights into all searches related to products, aiding in understanding product data from various perspectives.

	prod_id	total_cust_cnt
0	20175355001	13720297
1	20668578	4348230
2	20070132001	4242971
3	20812144001	3574510
4	20145621001	3102017
5	20007535001	3042138
6	21097012001	3035599
7	20527440	2934861
8	20671789	2920847
9	20028593001	2914054

Figure 2.7 Product bought by the most customers

	prod_category	total_cust_cnt
0	Vegetables	112446055
1	Fruit	76951642
2	Snacks	46568634
3	Natural Foods	44250995
4	In-Store	32576295
5	Cold Beverages	30781667
6	Milk & Eggs	30448371
7	Commercial	27714869
8	Meal Makers	25789370
9	Cheese/Butter/Margarine	23847899

Figure 2.8 Product categories bought the most customers

	prod_category	total_sales
1	Vegetables	409660753.33996785
2	Fruit	313889756.5899968
3	Natural Foods	290477157.06999964
4	Snacks	171215039.4400099
5	HMR	162000998.57000464
6	In-Store	149947369.4200053
7	Milk & Eggs	129179428.51000328
8	Cheese/Butter/Margarine	125803278.46000172
9	Cold Beverages	123416180.96000127
10	Dispensing	122376142.69999997

Figure 2.9 Product categories with the highest revenue

	prod_id	prod_category	total_sales
1	20733932	Dispensing	122113442.47999994
2	20027156	Lottery - Electronic	20566632.049999997
3	20175355001	Fruit	19308163.340000052
4	20252014	HMR	14932909.099999936
5	20425775001	Fruit	14282150.760000002
6	20159199001	Fruit	12245680.400000012
7	20188873	Milk & Eggs	11870592.579999987
8	21097012001	Fruit	11726112.26000002
9	20600985	Gourmet Foods	11454075.479999982
10	20049778001	Fruit	10783542.279999992

Figure 2.10 Products with the highest revenue

prod_id	total_sales	profit_margin	profit
20733932	122113442	0.30	36634032.60
20252014	14932909	0.30	4479872.70
20600985	11454075	0.35	4008926.25
20049221	4713834	0.75	3535375.50
20821992	10036564	0.30	3010969.20
21087193	10026234	0.30	3007870.20
2147483647	19308163	0.15	2896224.45
20188873	11870593	0.20	2374118.60
2147483647	14282151	0.15	2142322.65
2147483647	10234390	0.20	2046878.00

Figure 2.11 Products with the highest profits

prod_category	total_sales	profit_margin	profit
Natural Foods	290477157	0.40	116190862.80
Vegetables	409660753	0.20	81932150.60
Breakfast	80753430	0.75	60565072.50
Snacks	171215039	0.30	51364511.70
HMR	162000999	0.30	48600299.70
Fruit	313889757	0.15	47083463.55
Deli Cheese	116130075	0.35	40645526.25
Household Cleaning Needs	96248973	0.40	38499589.20
In-Store	149947369	0.25	37486842.25
Salad Bar	53187969	0.70	37231578.30

Figure 2.12 Product categories with the highest profits

Store Performance Rankings

	store_id	total_units_sold
0	1212	43983417
1	1050	40702229
2	1007	37965977
3	1004	37817869
4	1066	35972643
5	1021	33634193
6	1035	33423148
7	1027	30033180
8	1040	29387920
9	1188	29129309

Figure 3.1 Store with the most units sold

	store_id	total_weights_sold
0	1212	3738058.14
1	1007	3109467.16
2	1050	2887113.21
3	1066	2822230.02
4	1004	2679962.55
5	1021	2470113.91
6	1035	2435590.92
7	1011	2390760.24
8	1019	2269733.63
9	1027	2203426.15

Figure 3.2 Store with the the most weight sold

	store_id	total_sales
0	1212	2.009681e+08
1	1050	1.830785e+08
2	1004	1.722616e+08
3	1007	1.664551e+08
4	1066	1.649147e+08
5	1021	1.569396e+08
6	1035	1.547424e+08
7	1027	1.347106e+08
8	1011	1.341409e+08
9	1040	1.319841e+08

Figure 3.3 Store with the most transactions amount

	store_id	total_trans
0	1212	36401849
1	1050	32947201
2	1007	31739296
3	1004	30794895
4	1066	28929273
5	1035	27152074
6	1021	27136825
7	1027	24694438
8	1040	24295495
9	1188	23359225

Figure 3.4 Store with the most transaction

	store_id	total_cust_cnt
0	1212	36401852
1	1050	32947202
2	1007	31739310
3	1004	30794895
4	1066	28929273
5	1035	27152074
6	1021	27136825
7	1027	24694438
8	1040	24295495
9	1188	23359265

Figure 3.5 Store with the most customer visits

store_id	profit
1050	72482.77735
1212	60478.04310
1035	58522.26705
1007	54056.80620
1021	47282.40640
1051	46271.67130
1014	43862.71855
1019	43125.35065
1188	41467.93770
1032	41303.93950

Figure 3.6 Store with the most profit

The data presented in response to question three offers a comprehensive overview of key performance indicators across leading stores, encapsulated within a condensed dataset. It delineates a clear hierarchy in sales achievement, spanning from units sold to overall revenue generation, providing a snapshot of operational strengths and potential growth opportunities across the network. This analysis, distilled from a 0.1% data sample, is instrumental in guiding strategic decisions aimed at enhancing store performance and optimizing customer engagement.

Customer Segmentation

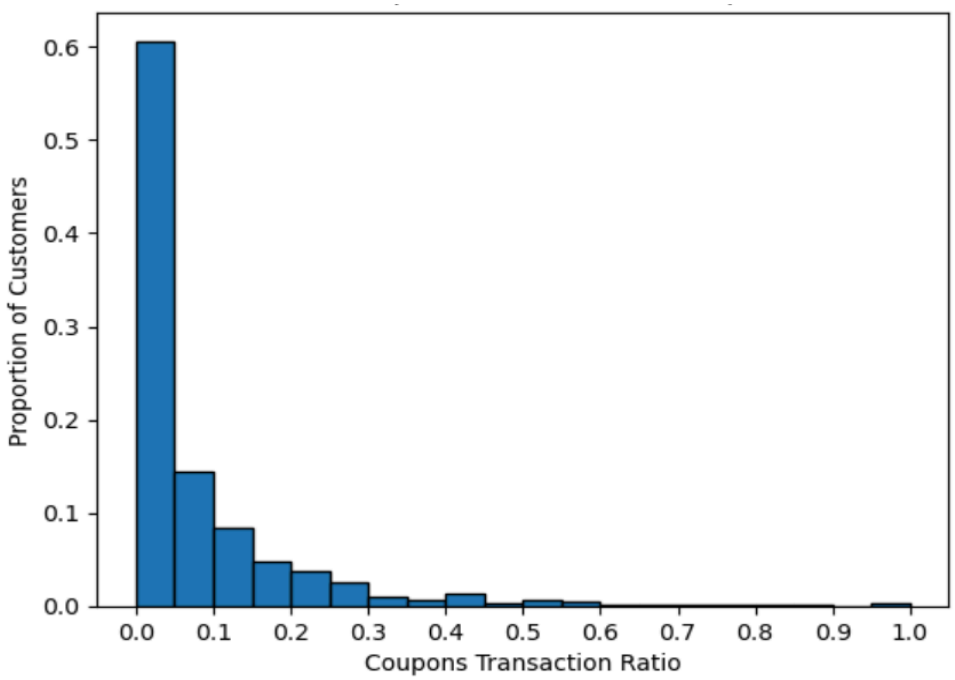


Figure 4.1 Distribution of coupons transaction ratio per customer

	cust_id	coupon_transactions	total_transactions	coupon_usage_frequency
0	1049278927	32	32	1.000000
1	33219887055	12	14	0.857143
2	1127456503	21	35	0.600000
3	60002977508460	21	42	0.500000
4	33220537979	26	60	0.433333
...	...	...	...	...
105	1148383776	3	97	0.030928
106	1124648385	10	326	0.030675
107	1125544177	11	361	0.030471
108	1019078208	1	33	0.030303
109	1055983674	1	33	0.030303

110 rows × 4 columns

Figure 4.2 Customer coupon usage summary

The histogram (Figure 4.1) displays the distribution of customers based on the ratio of their transactions in the "Coupons" category to their total number of transactions. The x-axis represents the ratio of transactions that are for coupons, divided into bins of 5% increments, ranging from 0% to 100%. The y-axis shows the proportion of customers falling into each of these ratio bins. Customers with a "Coupons" transaction ratio greater than 30% (Figure 4.2) are classified as "cherry-pickers." These are customers who predominantly engage in transactions that involve coupons, indicating a high propensity to take advantage of special offers and discounts provided by the retailer.

	Category	Customer Count Proportion	Profit Proportion
0	Never Use Coupon	0.484003	0.146583
1	Buy 1-3 Categories	0.013765	0.003421
2	Buy 50+ Categories	0.083333	0.451492

Figure 4.3 Certain patterns with .1% of the data

	Category	Customer Count Proportion	Profit Proportion
0	Never Use Coupon	0.489452	0.142310
1	Buy 1-3 Categories	0.012824	0.002219
2	Buy 50+ Categories	0.075960	0.443786

Figure 4.4 Certain patterns with .5% of the data

Figures 4.3 and 4.4 explore whether categorizing customers into certain groups could better illuminate their impact on profit margins. Since these analyses were conducted using sampled data, there was a concerted effort to mitigate bias arising from an overly small sample size. To this end, calculations were also performed with 0.05% of the data, showing no significant difference between the two sets of results (p-value >> 0.05) and distribution(Appendix III). From the graphs, we observe that customers using coupons contribute relatively less to total profits (49% of the population contributes 14.65% of total profits). Meanwhile, customers who purchase only specific items (with total transactions >10) make up 1% of the overall customer base yet only contribute 0.2% to profits, indicating their relatively lower value. In contrast, those who purchase a wide variety of categories (50+), making up 7.6% of the customer base, contribute a staggering 44.3% of total profits. This highlights how certain customer segmentation can stratify customer value, allowing for the tailored development of marketing and promotional strategies.

Strategic Product Groupings

prod_id	sum_profit
20733932	45596.6580
20252014	4913.9130
20600985	3655.7395
20049221	3573.2325
21087193	3280.5660
20175355001	3081.2475
20821992	2464.8570
20188873	2444.2780
20145621001	2284.0980
20812144001	2278.6440

Figure 5.1 Product with the most profit

prod_category	sum_profit
Natural Foods	108059.8400
Vegetables	80817.7580
Breakfast	59983.6500
Snacks	51193.7130
HMR	50811.4140
Fruit	48987.3285
Dispensing	45722.7360
In-Store	40809.8775
Deli Cheese	38441.6480
Cold Beverages	37022.5590

Figure 5.2 Category with the most sum profit

prod_id	sum_sales_amt
20733932	151988.86
20175355001	20541.65
20027156	17210.00
20252014	16379.71
20159199001	14541.00
20425775001	13442.76
20188873	12221.39
20128938001	11777.62
20049778001	11561.18
20145621001	11420.49

Figure 5.3 Product with the most sum sales amount

	prod_id	std_sales_qty	promotion_frequency	prod_category
0	20189092	393.022365	Often Promoted	Household
1	20745693	154.560021	Often Promoted	Garden
2	20131170001	153.833565	Often Promoted	Vegetables
3	20055266001	120.332856	Often Promoted	Fruit
4	20175355001	91.055438	Often Promoted	Fruit
5	20054039	86.163362	Often Promoted	Photo Image
6	20159690001	63.094072	Often Promoted	Fruit
7	21065203	56.976603	Often Promoted	Hardware/Automotive
8	20593619	56.438757	Often Promoted	Photo Image
9	20028593001	48.4934	Often Promoted	Fruit
10	20070132001	46.791668	Often Promoted	Vegetables
11	21097012001	46.228707	Often Promoted	Fruit
12	20755477	45.845392	Often Promoted	Garden
13	20826568001	45.456939	Often Promoted	Fruit
14	20639926	43.2869	Often Promoted	Cheese/Butter/Margarine
15	20128938001	41.81208	Often Promoted	Fruit
16	20080137001	41.454543	Often Promoted	Fruit
17	20040489001	41.051889	Often Promoted	Fruit
18	20049778001	40.841896	Often Promoted	Fruit
19	20095158001	37.97267	Often Promoted	Fruit

Figure 5.4 Promoted products top20 (chose the price variance above 75% quartile products)

KVI (Key Value Items) and KVC (Key Value Categories) are identified by calculating their total profit, highlighting the top 10 products and categories that contribute most significantly to the retailer's overall profit. Traffic-driving products are determined based on sales quantity, focusing on the top 10 products that attract the most customers. As for products that are often promoted, by analyzing the standard deviation of monthly sales quantities and using the 75th percentile of this variability as the threshold for "often promoted," we can identify the top 20 products with significant sales fluctuations, likely due to frequent participation in promotional activities. This method utilizes sales variability as an indirect indicator of promotional frequency, assisting retailers in understanding which products are more likely to be affected by promotional activities.

Store Grouping

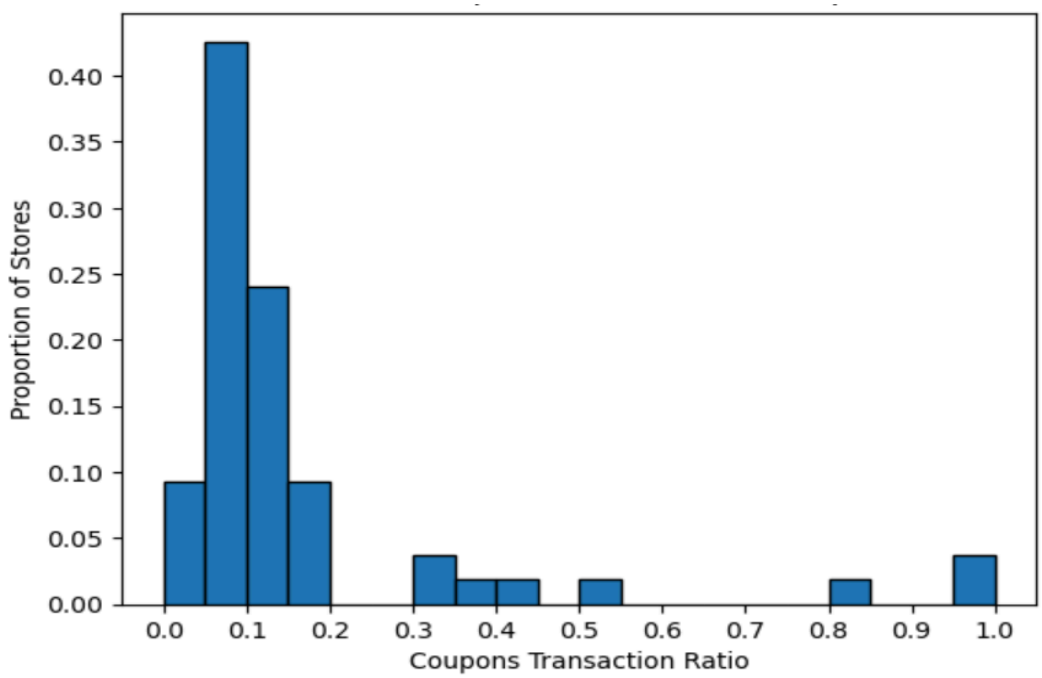


Figure 6.1 Distribution of coupons translation ratio per store

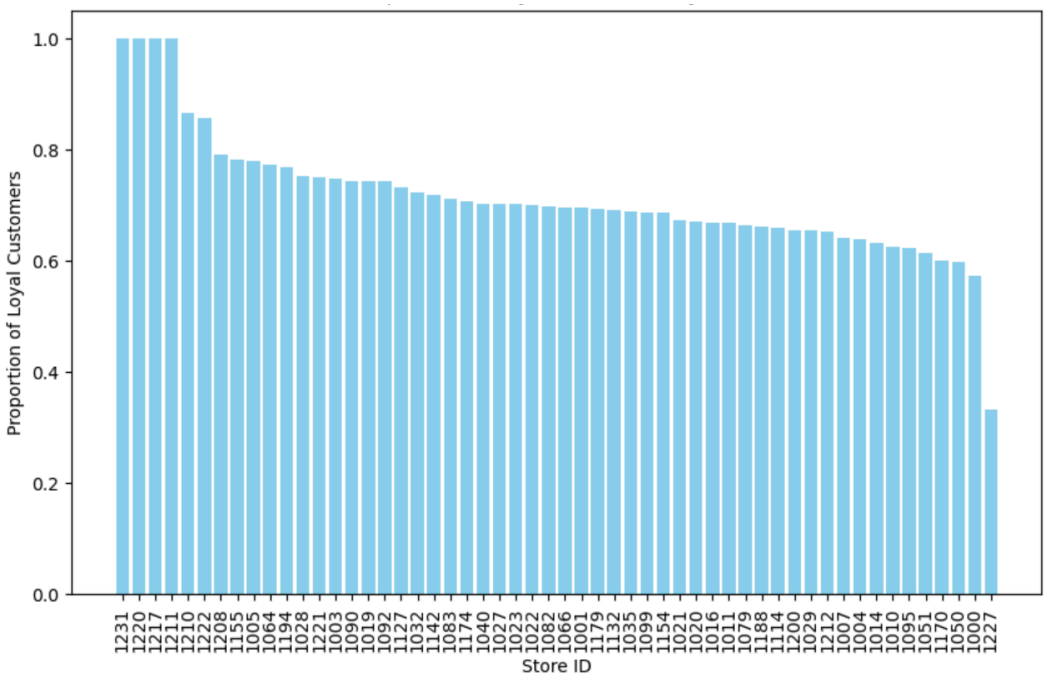


Figure 6.2 Proportion of loyal customers by store

store_id	coupons_ratio
1227	1.000000
1231	1.000000
1217	0.833333
1210	0.529412
1211	0.400000
1221	0.357143
1220	0.333333
1222	0.333333

Figure 6.3 Stores with the most coupons usage ratio

	store_id	loyal_customer_proportion
53	1231	1.000000
49	1220	1.000000
48	1217	1.000000
46	1211	1.000000
45	1210	0.866667
51	1222	0.857143
44	1208	0.791667
37	1155	0.781818
4	1005	0.780822
23	1064	0.773585

Figure 6.4 Stores with the most loyal customer proportion

Figure 6.1 illustrates the variance in stores' coupon transaction ratios, distinguishing between "cherry-pickers" (high ratio) and "regularly visited" stores (low ratio), highlighting different customer attraction strategies. Focusing on stores with a ratio above 0.3, which constitutes 14.81% of the total, is rational. This segment likely leverages coupons effectively for customer engagement, reflecting on their marketing success and customer discount affinity. However, a high reliance on coupons could impact profitability and loyalty.

In our report, we define loyal customers as those with a cust_id of exactly 10 digits. We calculate the loyalty metric by dividing the number of unique loyal customer visits by the total unique visits per store. The bar chart illustrates the top 10 stores by this metric, indicating a high level of engagement from loyal customers. This concise analysis highlights stores with the strongest customer loyalty, useful for informed marketing and retention strategies.

Figures 6.3 and 6.4 further dissect the nuances of customer engagement through coupon usage and loyalty metrics. Figure 6.3 shows a concentrated look at the stores with the highest coupon transaction ratios, potentially pointing to strategic discount use to boost short-term traffic and sales. The focus on stores with ratios above 0.3 suggests an effective use of coupons as a tool for driving customer visits, which might indicate both an opportunity in terms of attracting price-sensitive segments and a risk if discounts undermine long-term profitability. Figure 6.4 delves into the proportions of loyal customers at each store, revealing where the most dedicated shoppers are concentrated. The stores with the highest proportion of loyal customers demonstrate the successful fostering of repeat business, likely through exceptional service, product quality, or customer satisfaction.

Text Mining

Objective

We took on the task of making sense of our product descriptions through text mining. In short, we improve the accuracy of product categorization by identifying mismatches using text mining techniques, specifically TF-IDF vectorization, and optimizing the number of features to enhance model performance.

Process Overview

- **Mismatch Identification**: A function, check_for_mismatches_v2, was developed to identify mismatches between product descriptions and their assigned categories. The function iterates through the dataset, checking if product descriptions contain category-specific keywords. Products lacking these keywords are flagged as potential mismatches.

- **Text Preprocessing**: Product descriptions were preprocessed to improve the relevance of the TF-IDF features. The preprocessing steps included converting text to lowercase, removing punctuation, and excluding stopwords. To see the big picture of our findings, we also created a word cloud, which visually showed us which terms were most frequent.

- **TF-IDF Vectorization**: The TfidfVectorizer was utilized to extract features from the cleaned product descriptions, initially setting max_features to 500. This step aimed to identify the most significant words for each product category.

- **Keyword Enhancement**: Based on the TF-IDF results, the existing category-specific keyword lists were updated to include new significant terms identified through TF-IDF, enriching the keyword repository for better mismatch identification.

- **Model Training and Evaluation**: A series of experiments were conducted to find the optimal number of TF-IDF features. A range of max_features values was tested, and the performance of a MultinomialNB classifier was evaluated using accuracy, precision, recall, and F1 score as metrics.

- **Optimization Results**: The experiments indicated that increasing the number of features generally led to better accuracy. However, concerns about potential false positives (items incorrectly categorized due to the model's training) were raised.

- **Feature Selection**: To address potential overfitting and the false positive issue, an optimal number of features was sought through performance metrics evaluation, considering precision, recall, and F1 score across a defined feature space (Appendix IV).

- **Final Application**: With the optimal number of features identified (set to 5000 for demonstration), the TfidfVectorizer was reapplied to extract refined keywords for each category. The mismatch identification function was run again to detect mismatches with the enhanced keyword set.
- **Mismatch Handling**: Products identified as mismatches in crucial categories ("Reading" and "Colour Cosmetics - Mass") were filtered out due to the uniqueness of their description for further analysis, ensuring a focused approach to addressing categorization issues.

Results

Expanding on our results, we found that enhancing the feature set for natural language processing from 500 to 5000 TF-IDF features significantly reduced mismatches in product categorization from 2235 to 163 rows. This allowed us to manageably sift through the data manually, updating our dataset for accuracy and future utility.

While this refinement in features achieved a notable reduction in mismatches, it introduced a complexity that merits discussion. The increased false positives highlight a delicate balance in data mining—between detailed feature representation and the risk of overfitting or misclassification. It was during manual review that certain nuances came to light, such as rare instances of non-standard spellings or hyphenated words, that algorithms misidentified as mismatches (Appendix V). These nuances underscore the limitations of relying solely on algorithmic processing for data quality and categorization.

Further, these findings underscore the need for caution in interpreting algorithmic outputs. While the mismatches currently seem manageable and their rectification appears not to impact our future modeling efforts significantly, overlooking even minor inconsistencies could lead to larger issues. For instance, in the case of predictive modeling, these small errors could potentially cascade, affecting the reliability of customer behavior predictions and the efficacy of the recommendation system.

Therefore, while the impact of mismatches on our models may appear minimal at present, continuous improvement of data preprocessing techniques is crucial. This includes developing more sophisticated algorithms that can recognize and adapt to the nuances of human language and input variations, thereby reducing the need for manual intervention. Ensuring data accuracy is not a one-time task but an ongoing process that directly contributes to the strength and reliability of our analytical models. Our discussions and decisions around these data challenges set the stage for more advanced analytics and more accurate, data-driven decision-making at ACSE Supermarket.

Conclusion

Summary

In conducting our analysis of ACSE Supermarket's extensive customer data, we embarked on a detailed journey through the landscape of purchasing behaviors, seeking actionable insights for strategic planning and model development. Our balanced approach involved a rigorous yet practical data preparation phase, where we meticulously cleaned and refined the dataset to ensure accuracy without losing sight of meaningful patterns and customer interactions.

We faced the challenge of managing a vast amount of data making critical decisions on how to effectively trim and filter the dataset to highlight relevant customer behaviors. This included addressing outliers and ensuring that the data reflected genuine transactions, a task that required both precision and a keen understanding of retail dynamics.

Our exploration into customer purchase frequencies led us to identify distinct customer segments, revealing valuable insights into how different groups contribute to the supermarket's profitability. Through careful segmentation, we uncovered patterns that pointed to the potential for more focused marketing strategies and the optimization of product offerings to enhance customer engagement and loyalty.

Summarizing the intricacies of our findings, we noted the importance of certain customer behaviors in driving profit margins. For instance, our analysis highlighted the disproportionate impact of loyal customers who purchase a wide range of products, underscoring the need for targeted approaches to foster and capitalize on this loyalty.

The insights gleaned from our comprehensive analysis are set to inform future model building and deeper analytical efforts. By sifting through the details and understanding the nuances of customer purchasing behavior, we've laid the groundwork for ACSE Supermarket to tailor its strategies more effectively, ensuring that future endeavors are both data-driven and aligned with customer needs. Our methodical yet accessible approach aims to equip ACSE Supermarket with the knowledge to navigate the competitive retail landscape successfully, leveraging customer insights to drive profitability and customer satisfaction.

Recommendations

Based on our comprehensive analysis of ACSE Supermarket's customer data, we recommend a series of targeted strategies to enhance profitability and customer engagement:

- **Leverage Customer Segmentation**: Our findings highlight the value of different customer segments to ACSE's profitability. We recommend tailoring marketing and

promotional strategies to these segments, particularly focusing on loyal customers who purchase a wide range of products. Personalized marketing campaigns and loyalty programs can help deepen these relationships, encouraging more frequent visits and larger basket sizes.

- **Optimize Product Assortment**: Given the significant contribution of certain product categories to overall profits, ACSE should consider optimizing its product assortment. This involves analyzing sales data to identify high-performing categories and products, and adjusting inventory levels accordingly to meet customer demand without overstocking less popular items.
- **Enhance Customer Experience**: To attract and retain valuable customer segments, improving the overall shopping experience is key. This could include streamlining the checkout process, enhancing in-store navigation, and investing in customer service training. Additionally, leveraging technology to offer personalized shopping recommendations and promotions can further enhance the customer experience.
- **Data-Driven Decision Making**: Continue to invest in data analytics capabilities to refine understanding of customer behaviors and preferences. Ongoing analysis of transactional data can uncover emerging trends and allow for quick adjustments to marketing and operational strategies.
- **Address New and Low-Value Customers**: While focusing on high-value customers, ACSE should not neglect newer or currently low-value customers. Implementing strategies to engage these groups, such as introductory offers or targeted communications, can help convert them into more profitable segments over time. It’s also important to analyze the reasons behind low spending to identify any barriers and address them effectively.
- **Review and Adjust Data Trimming Practices**: Given the impact of our data preparation methods on the insights obtained, ACSE should periodically review and adjust these practices as needed. This ensures that the dataset remains representative and that the analysis captures the full spectrum of customer behavior.

Implementing these recommendations will enable ACSE Supermarket to leverage our analysis insights, boosting profitability and customer satisfaction. Tailored strategies for customer segments, optimized product offerings, and data-driven decision-making are key to strengthening ACSE's market position.

Appendices

I: Most Frequent Visitors

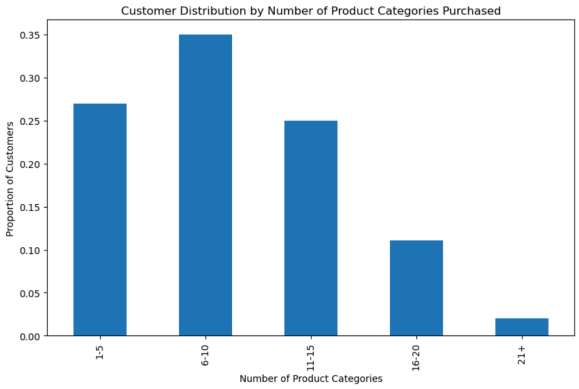
cust_id	frequency
1135252713	90265
1147458804	71181
1143554806	63729
1130048984	43183
1127306013	43004
1045022989	37806
1133211531	32138
1130576377	31858
1130724383	31041
1150823229	29417

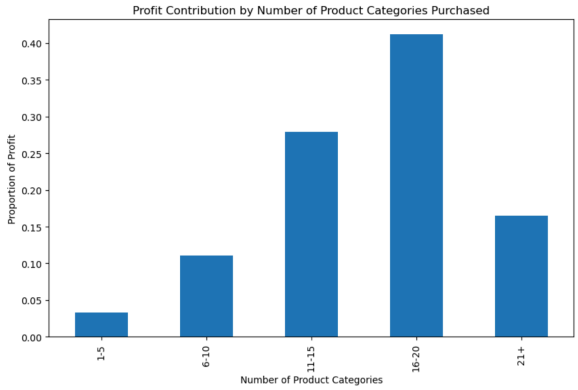
II: Loyal customers ration on profit

	Metric	Value
0	Loyal Customer Ratio	0.598602
1	Loyal Customer Profit Ratio	0.840934

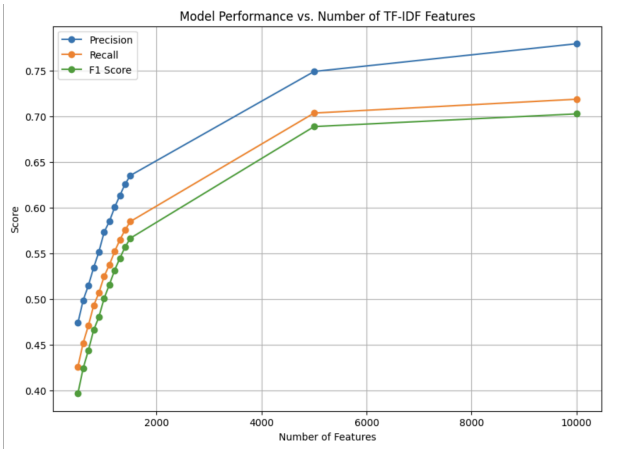
III: Distribution

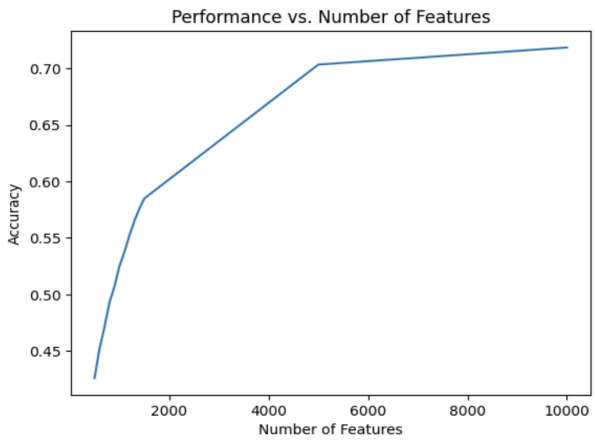
	Category	Z-Score	P-Value
0	Never Use Coupon	-0.243773	0.807407
1	Buy 1-3 Categories	0.183715	0.854237
2	Buy 50+ Categories	0.608932	0.542570





IV: Text mining feature selection

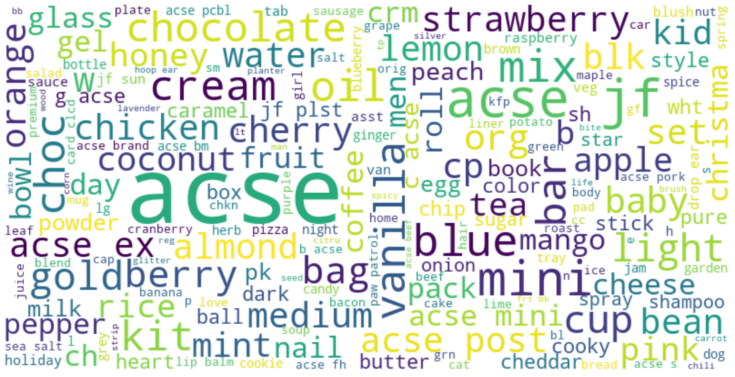




V: Text mining result examples

	Product ID	Description	Category
0	20005051	ST-ALBERT MOZZ.PART.SKIM	Deli Cheese
4	20017627	10'X30'" BANNER"	Photo Image
5	20017733	8'X24'" BANNER"	Photo Image
6	20018002	ST-AUBIN	Deli Cheese
7	20018648	S&B CHOC-O-MACS	Bulk Foods
...	...	...	...
454	20313741002	COCA-COLA	Cold Beverages
455	20316026002	COCA-COLA	Cold Beverages
456	20337420002	CHING-INSTANT NOODLES-SCWN	Canned
457	20931457001	ENER-C TANGERINE-GRAPEFRUIT	Natural Foods
458	20962641001	NASEBERRY/CHIKOO	Fruit

163 rows × 3 columns



References

OpenAI. (2024, March 7). Information on profit margins across various retail categories [ChatGPT interaction]. ChatGPT.

Bean-Mellinger, B. (2018, November 8). What Is the Typical Gross Margin for Beauty Products? Bizfluent.https://bizfluent.com/info-12211951-typical-gross-margin-beauty-products.html

What is the Average Profit Margin of the Cosmetic Industry? (n.d.). Freelance Formulations. https://www.freelanceformulations.com/post/what-is-the-average-profit-margin-of-the-cosmetic-industry

Kinnier, A. (2022, August 9). I've analyzed the profit margins of 30,000 gas stations. Here's the proof fuel retailers are not to blame for high gas prices. Yahoo! Finance. https://finance.yahoo.com/news/ve-analyzed-profit-margins-30-134200907.html

Hoeksema, A. (2023, April 24). 9 Gasoline/Gas Stations Industry Financial Statistics: Sales, Expenses, Profit and More. ProjectionHub. https://www.projectionhub.com/post/gas-stations-industry-financial-statistics

Is Skin Care Business Profitable in 2022? (2022, June 27). H&H Australia. https://www.hhaustralia.com.au/the-business-of-skincare/is-skin-care-business-profitable/

How to Start a Profitable Skincare Business in 2024. (n.d.). Step By Step Business. https://stepbystepbusiness.com/business-ideas/start-a-skincare-business/

Skincare Market Size, Share & COVID-19 Impact Analysis. (n.d.). Fortune Business Insights. https://www.fortunebusinessinsights.com/skin-care-market-102544

The 2022 Medical Spa State of the Industry Executive Summary in Context. (n.d.). American Med Spa Association.