`

ACSE Supermarket Data Understanding

AI&ML at Scale

2024 March 08

Anne Lim, Benny Uhoranishema, Silvia Lee, Yaeeun Lee, Zhuhuang Jiang

Executive Summary

This report outlines the initial analysis and findings conducted by our analytics consulting team on the transactional and product data provided by ACSE Supermarket. Our objective is to leverage these insights to develop a robust recommender system that will support ACSE in various operational aspects: supply chain management, store operations, supplier relations, and marketing strategies. Through our comprehensive data analysis, we aim to demonstrate a deep understanding of ACSE's business, identifying key customer segments, high-performing products and stores, and addressing data quality issues to ensure the successful implementation of the recommender system. Given the complexity and volume of data involved, our team employed a systematic approach to data cleaning, sampling, and analysis, ensuring that the insights generated are both accurate and actionable.

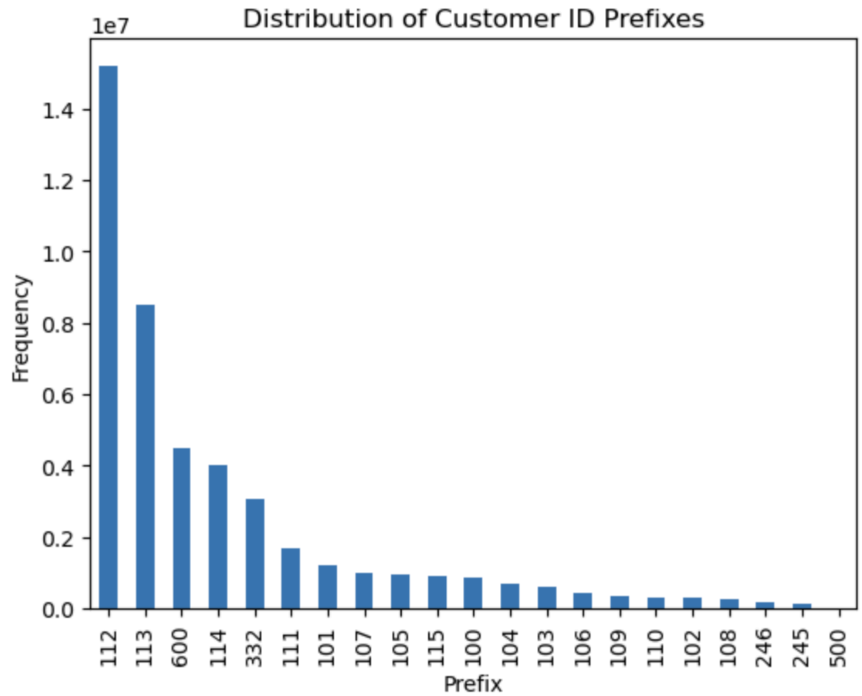
Data Understanding

1. Customer
    a. Loyal Customers
        i. Our team found out that customer_id had certain patterns and assumed customer loyalty was reflected into this pattern. After printing out the length of customer ids, our team found our that customer ids have 10-14 digits with 10 digits being the most frequent. Our team eyeballed the first 3 digits having a pattern. Thus, we extracted the first 3 and last 3 characters and made 2 separate lists: prefixes and suffixes.

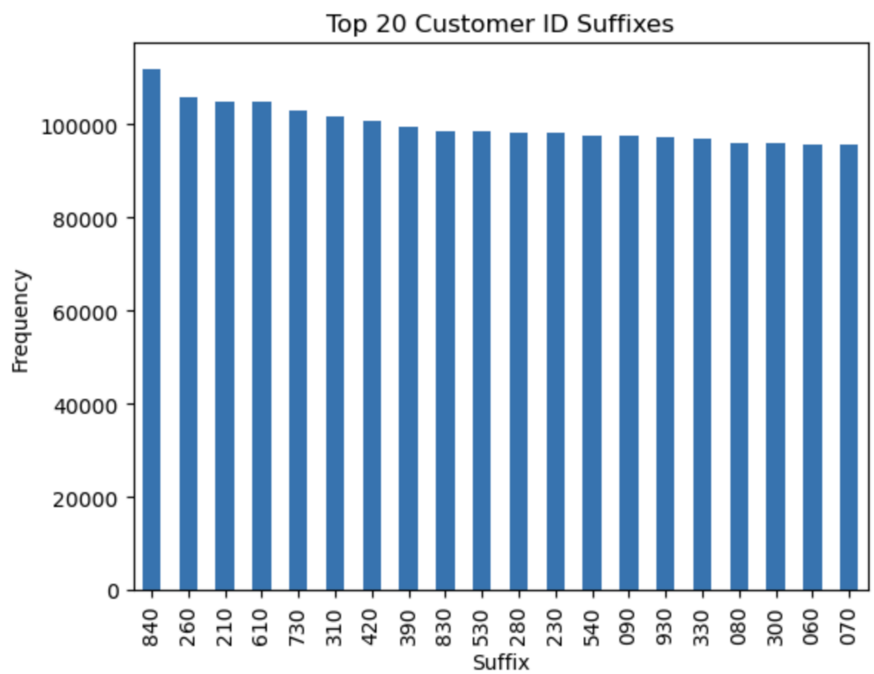
1.
```
cust_id_length
10    37257598
14     4790055
11     3070819
Name: count, dtype: int64
cust_id_prefix
112    15183129
113     8516965
600     4498285
114     4004910
332     3070819
111     1702753
101     1229680
107     1000803
105      936767
115      893144
100      857116
104      696106
103      588811
106      423438
109      345710
110      320537
102      308328
108      249401
246      167310
245      113324
500       11136
Name: count, dtype: int64
```

2.
```
cust_id_suffix
840    111887
260    105743
210    104965
610    104839
730    103015
        ...
725     23765
079     23498
212     23488
456     20687
857     20396
Name: count, Length: 1000, dtype: int64
```

ii. We wanted to see the distribution of what is the most repeated prefixes from the customer ids and created bar plots. Customer ids starting with 112 are substantially frequent compared to other prefixes. For suffixes, the pattern was indistinguishable, so we narrowed down our scope to top 20 frequent suffixes, where we see that suffixes like ‘840’ is the most repeated. Overall, prefixes are more important to consider than suffixes to analyze further if customer loyalty is relevant to patterns in customer ids.



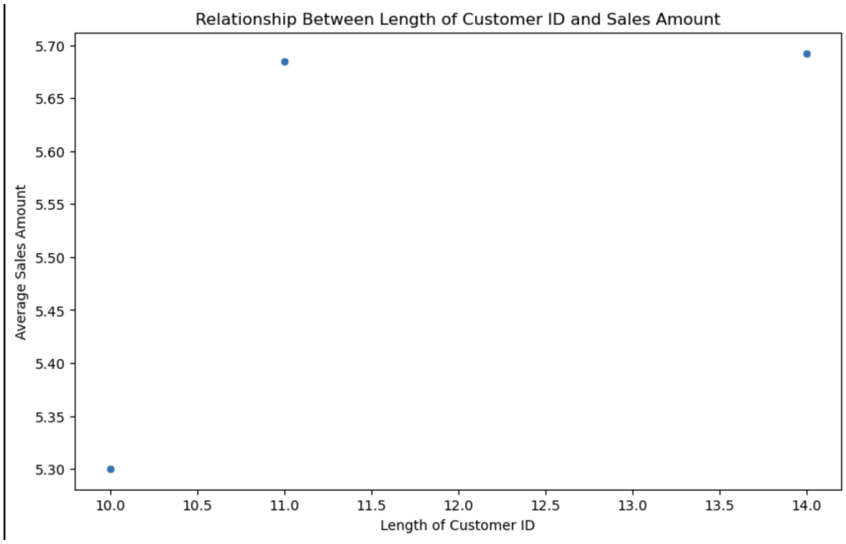
1.

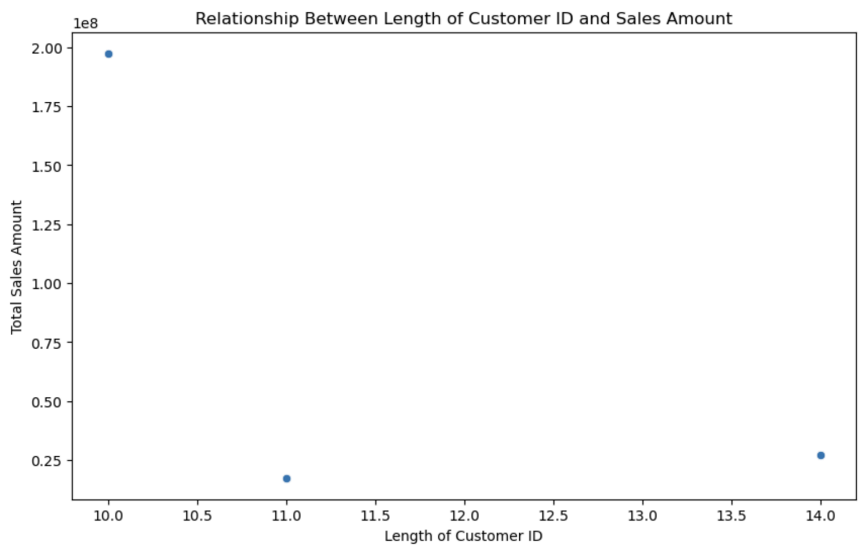


2.

iii. After identifying patterns in prefix/suffix of customer ids, we wanted to see their correlation between other features like store id, sales quantity, or samles amount to see if "royal" customers' behaviors influence customer id having a certain pattern. As a result, the length of a customer id does not influence how much a customer spends on average. However, length of customer id does influence how much a customer spends on total. This indicates that the store would care more on total spending relative to average spending in order to segment “loyal” customers.

iv.

1. 

2. 

v. Our team decided to conduct 3 statistical analysis methods to identify the correlation between customer ID prefixes and store ID, sales amount, or sales quantity. Based on the distribution of different features, we decided to use chi-square to identify the correlation between store id and customer ID prefixes, ANOVA for sales quantity and customer ID prefixes, and spearman for sales amount and customer ID prefixes.

1. Given the chi-square statistic is 1793418 and the p-value is 0, we can reject the null hypothesis that there is no association between the two variables in the population. This shows that certain stores may have a distinct distribution of customer ID prefixes, which might reflect loyal customers’s prefixes starting as 112. There’s a

statistically significant association between sales quantity and customer id's prefixes as well. The F-statistic is 82.4 and the p-value is 0. The p-value is greater than 0.05, so the correlation between the customer id prefix and sales amount is not statistically significant.

2.
```
Chi-square test result: Chi2 = 1793418.4746711906, p-value = 0.0
ANOVA test result: F = 82.40298276424394, p-value = 0.0
Spearman correlation: r = -0.0008746612328660701, p-value = 4.2252338584362164e-09
```

vi. Now that we identified "loyal customers", we want to identify loyal customers' top 10 purchased products and top 10 frequently visited stores. Assuming that ACSE's plastic bags and ACSE Plus Points aren't considered a "purchase product" due to being a product or service offered during checkouts, we see that most loyal customers purchase grocery items such as vegetables and fruits as well as frequently visit store id of 1212, 1007, 1050, and etc.

1.

Top 10 Products among Loyal Customers:

	Product	Frequency
0	ACSE PLASTIC BAGS	973213
1	BANANA	446408
2	ACSE GREEN PC POINTS	154364
3	PENNY ROUNDING - DO NOT TOUCH	141243
4	CUCUMBER ENGLISH	137316
5	ACSE PLUS POINTS	130552
6	ACSE GRADE A EGGS LARGE WHITE, EA	114397
7	PEPPERS RED SWEET	102749
8	BROCCOLI	100337
9	COLL DISC PROG DISC	97833

2.

	Product	Frequency
0	Grocery	7306756
1	Produce	6220170
2	Dairy	2475034
3	Meat	1425472
4	Natural Foods	1394560
5	Home	1314203
6	Deli	1268596
7	Frozen	1087916
8	Bakery Instore	1014678
9	Bakery Commercial	868325

3.

	Product	Frequency
0	Front End Bags	973482
1	Root Veg	912894
2	Milk	653485
3	Field Veg	648785
4	Salty Snacks	613889
5	Cheese	590403
6	Berries/Cherries	532338
7	Bananas	525162
8	BASKET COUPONS	517652
9	Cooking Veg	516142

4.

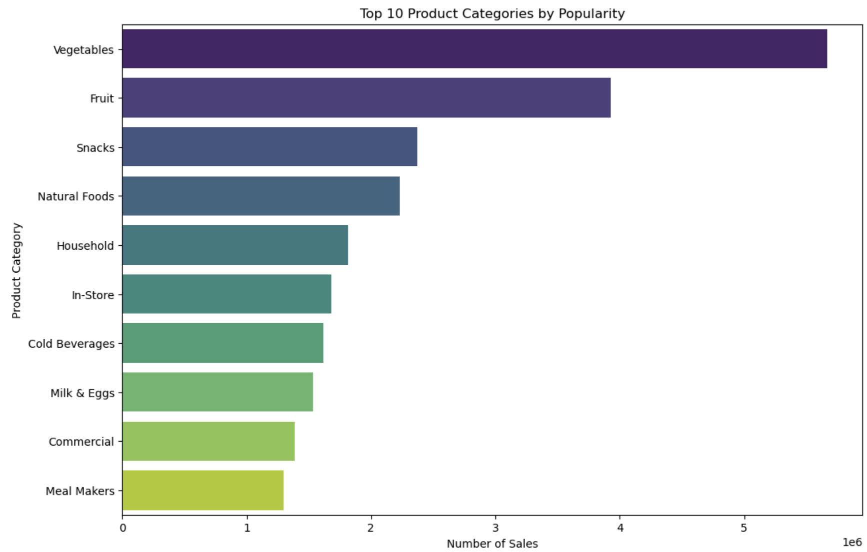
Top 10 Stores among Loyal Customers:

	Store ID	Frequency
0	1212	1252194
1	1007	1120233
2	1050	1035407
3	1004	1018961
4	1035	976257
5	1066	965145
6	1021	964943
7	1027	845607
8	1188	819181
9	1040	778377

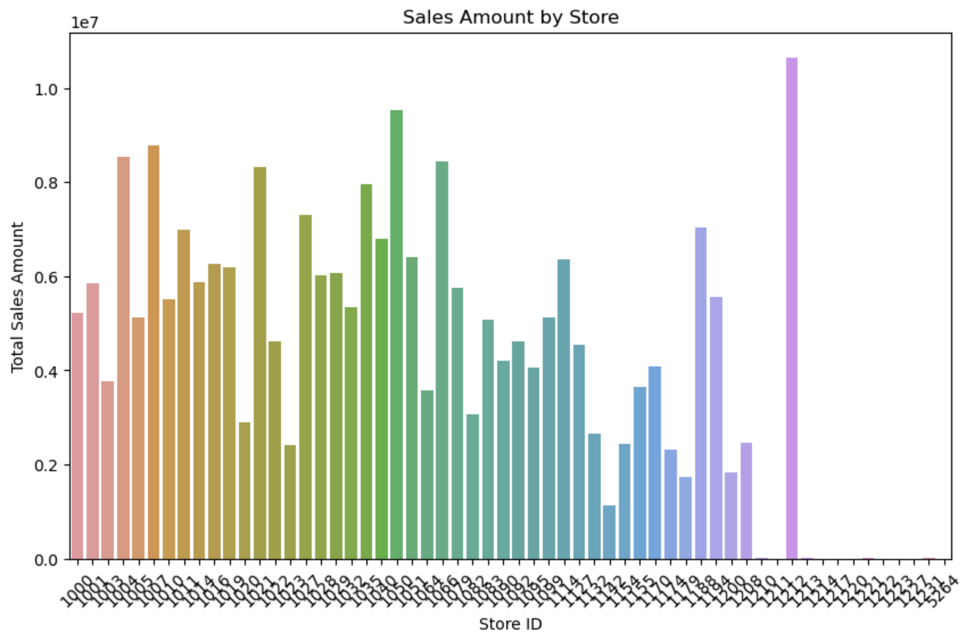
b. Target Customers

i. Our team assumed that the target customers were those who purchased items more than once. Thus, we wanted to look at the top 10 products that are being sold and the store with the highest sales to narrow down the scope. Again, we see that vegetables, fruit, and snacks have the highest sales amount as well as most customers visiting store id 1212. Now we want to see the percentage of customers who purchased those products.

1. % of customers who regularly buy vegetables: 40.13%
2. % of customers who regularly buy fruits: 36.75%
3. % of customers who regularly buy snacks: 32.77%



4.



c. Best Customers

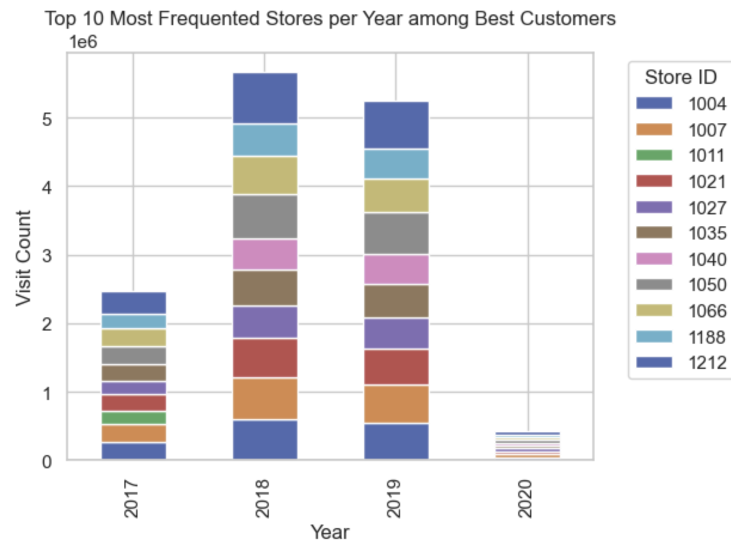
i. Our team used the RFM model to identify ACSE's best customers. We have tried to aggregate the best customers on a yearly basis to identify further changes in trends over the 4 years.

1. Recency - It seems that the customer recently visited the store almost 3 weeks or 1 month before our most recent date '2020-01-31'. Most % of the best customers focused in the last 3 weeks or 1 month probably indicates either customers' biweekly shopping patterns or issues with data being collected recently.

a.
```
Percentage of customers visited within the last 1 week: 7.12%
Percentage of customers visited within the last 2 weeks: 10.24%
Percentage of customers visited within the last 3 weeks: 12.49%
Percentage of customers visited within the last 1 month: 14.71%
```

2. Frequency -

a.



b.
```
Top 10 Most Frequently Visited Stores among Best Customers:
 Store ID  Visit Count
     1212      1736719
     1050      1485427
     1007      1378900
     1004      1356663
     1021      1318075
     1066      1251388
     1035      1233832
     1188      1088702
     1027      1050114
     1040      1023937
```

3. Monetary - Best customers spend around a yearly average of $241-$245. However, ACSE's average spending amount is $1261-$1327. From the minimal change in spending from both the customers and the store within a year, we can see that ACSE's customers are more engaged or have a higher reliance on the stores than the average other customer-grocery store relationships in terms of spending.

a.

```
Yearly Average Number of Visits for Best Customers:
 year  Yearly_Avg_Visits
 2017         132.183907
 2018         245.863799
 2019         241.877936
 2020          35.804044

Yearly Average Spending Amount for Best Customers:
 year  Yearly_Avg_Spending
 2017           695.992142
 2018          1261.968796
 2019          1327.856301
 2020           194.246864
```

2. <u>Product</u>

1. Summary Statistics
   a.

	trans_id	store_id	cust_id	prod_id	sales_amt	sales_qty	sales_wgt	prod_unit_qty_count	prod_uom_value
count	4.511847e+07	4.511847e+07	4.511847e+07	4.511847e+07	4.511847e+07	4.511847e+07	4.511847e+07	4.511847e+07	4.511847e+07
mean	1.833146e+17	1.068495e+03	6.150609e+12	7.676709e+09	5.368145e+00	1.257294e+00	1.000575e-01	1.585711e+00	2.175099e+02
std	7.761259e+15	7.263049e+01	1.804147e+13	9.850547e+09	1.722357e+01	1.087323e+00	3.723249e-01	7.654873e+00	2.593836e+02
min	1.706240e+17	1.000000e+03	1.000003e+09	2.000000e+07	-1.837000e+03	-2.890000e+02	-1.430000e+01	1.000000e+00	1.000000e-02
25%	1.803030e+17	1.016000e+03	1.124617e+09	2.055459e+07	2.490000e+00	1.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00
50%	1.810140e+17	1.040000e+03	1.129342e+09	2.101176e+07	3.990000e+00	1.000000e+00	0.000000e+00	1.000000e+00	1.200000e+02
75%	1.906020e+17	1.099000e+03	1.144111e+09	2.016295e+10	5.990000e+00	1.000000e+00	0.000000e+00	1.000000e+00	3.750000e+02
max	2.001310e+17	5.264000e+03	6.000314e+13	2.124746e+10	3.814250e+04	1.604000e+03	5.333300e+02	7.080000e+02	7.480000e+03

- The statistical summary of the dataset showcases over 45 million transactions, with a wide range of sales amounts and quantities, highlighting considerable variability in customer purchases and product sales across the supermarket chain.

2. Checking out number of rows and columns
   a.

```
Number of rows: 45118472
Number of columns: 17
```

- After sampling the transactions table and joining it with the product table, we have created a combined dataset that consists of **45,118,472** rows and **17** columns. Our EDA was basically conducted on this extensive, merged dataset to extract insights and understand purchasing patterns.

3. Checking out number of unique values
   a.

```
trans_id                5455900
trans_dt                    928
store_id                     58
cust_id                  404660
prod_id                  103540
sales_amt                 27878
sales_qty                   224
sales_wgt                  2181
prod_desc                 97997
prod_section                 32
prod_category               100
prod_subcategory            411
prod_type                  1918
prod_mfc_brand_cd          4272
prod_unit_qty_count          77
prod_count_uom               12
prod_uom_value             1392
dtype: int64
```

- The dataset contains a rich diversity of data with millions of unique transactions, customers, and products, along with a broad range of sales amounts and quantities, reflecting the extensive operations and varied inventory of the ACSE Supermarket chain.

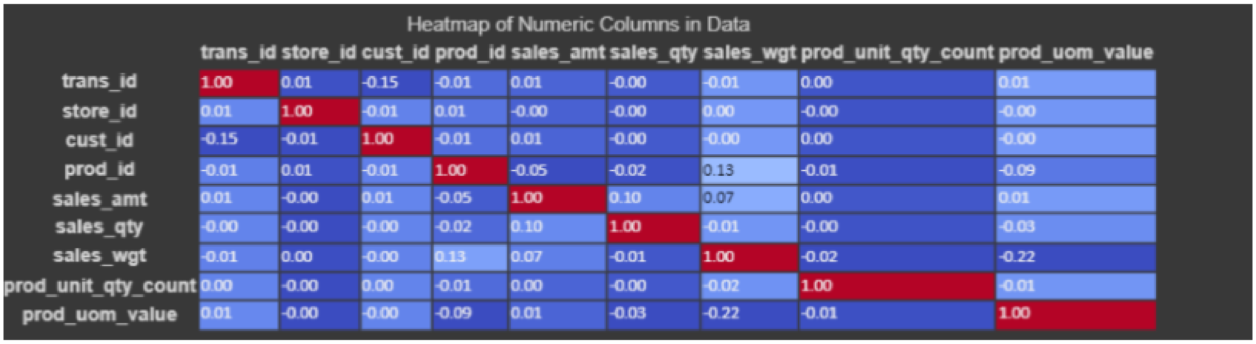
4. Checking Missing Values

   a.

```
trans_id                   0
trans_dt                   0
store_id                   0
cust_id                    0
prod_id                    0
sales_amt                  0
sales_qty                  0
sales_wgt                  0
prod_desc                  0
prod_section               0
prod_category              0
prod_subcategory           0
prod_type            1226068
prod_mfc_brand_cd          0
prod_unit_qty_count        0
prod_count_uom             0
prod_uom_value             0
dtype: int64
```

- The dataset is predominantly complete except for the 'prod_type' column, which has 122,068 missing entries. Nevertheless, since every product is already categorized into a subcategory with no missing data, this omission is unlikely to significantly impact the analysis, allowing us to proceed without concerns for data integrity at the product classification level.

5. Heatmap to show correlation between columns



- The heatmap of numerical columns in this dataset indicates mostly low correlation values between variables, suggesting that there is no strong linear relationship among these features, which could probably mean a diverse range of factors influence sales amounts and quantities within the ACSE Supermarket's transactions.

1. **A number of unique products have been purchased by customers.**

```
The number of unique products purchased by customers is: 103540
```

- The approach involved isolating the product hierarchy columns and removing duplicate entries to identify unique product combinations within the dataset.

- The analysis revealed a diversity in ACSE's product range, with 103,540 unique items purchased, suggesting a rich and varied product assortment available to customers.

- This conclusion is supported by the count of unique entries in the product hierarchy after duplicates were excluded, providing concrete data on the number of distinct products transacted.

- In terms of profitability, such a wide assortment of products could be indicative of ACSE's strategy to meet various customer needs, potentially leading to higher market penetration and customer retention. While this does not directly measure profitability, it is often associated with positive financial outcomes, as a larger product mix can attract and satisfy a broader customer demographic, potentially leading to increased sales volumes and, by extension, improved profit margins, assuming effective inventory and supply chain management.

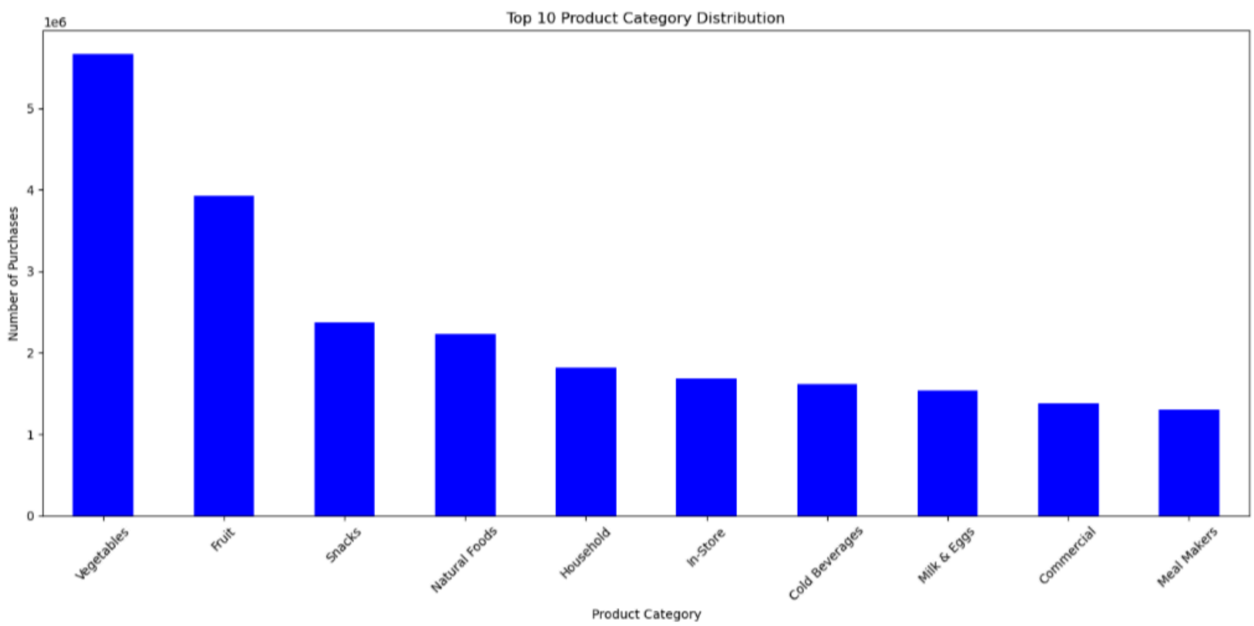
2. Product category distribution. Seeing which category comprises the majority of purchased products

a.

```
Product category distribution:
 prod_category
Vegetables                 5669242
Fruit                      3928211
Snacks                     2370717
Natural Foods              2233586
Household                  1813140
                            ...
Other                            5
Cosmetic Fragrances              3
Jewelery & Accessories           3
Supplies                         2
Cosmetic Treatments              1
Name: count, Length: 100, dtype: int64
```

- The top category with the majority of purchased products is: 'Vegetables' with 5669242 purchases.

b.



- The procedure involved calculating the frequency of product categories from the merged dataset and identifying the category with the highest number of purchases.

- The 'Vegetables' category is the most purchased, with a total of 5,669,242 purchases, reflecting consumer preference and potentially indicating a market trend towards food and health-conscious products.

- This is evidenced by the value counts for each category and the bar chart visualization that displays the distribution, clearly showing 'Vegetables' as the top category amongst the product range.

- In terms of profitability, the dominance of the 'Vegetables' category could suggest a steady demand and the possibility of consistent revenue from these products. Given that perishable goods like vegetables often have a higher turnover rate, this could imply a healthy cash flow for ACSE. However, it is important to note that while high-volume sales can be profitable, vegetables typically have lower margins compared to other categories, so the actual impact on profitability would also depend on effective inventory management and waste reduction strategies.

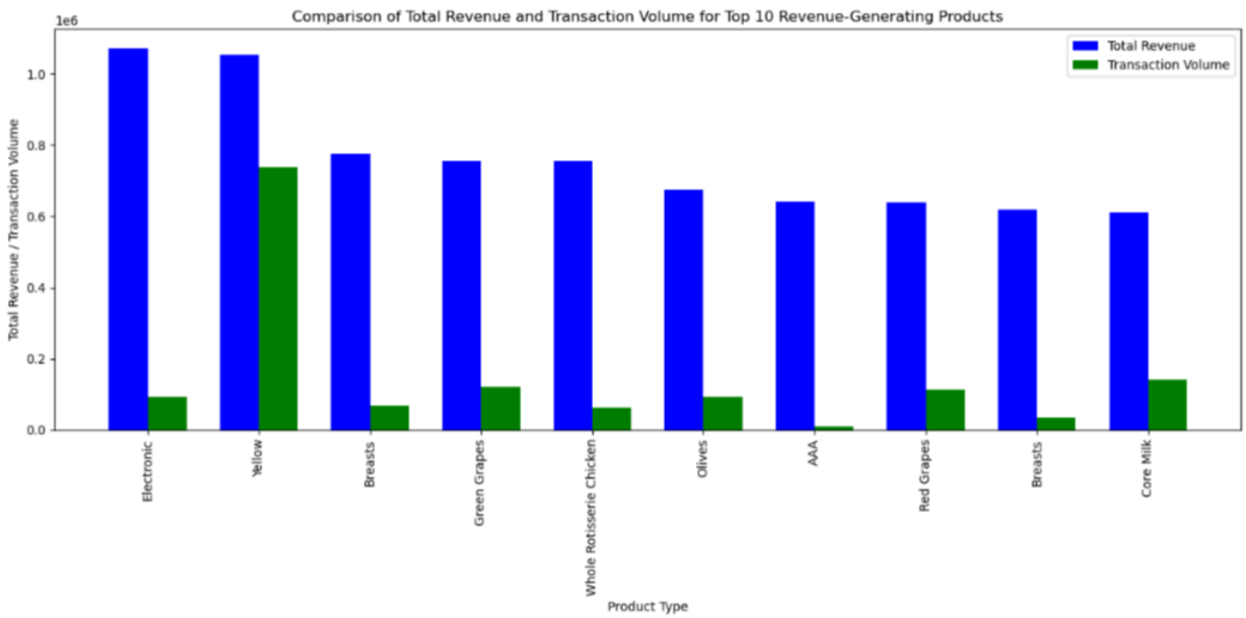
3. The products with the best revenue and transaction volume.
   ○ Show the product distribution by revenue

a.

Top 10 products by revenue:

	prod_id	prod_section	prod_category	prod_subcategory	prod_type	total_revenue	transaction_volume
1218	20027156	Customer Service	Lottery - Electronic	LOTTERY - ELECTRONIC	Electronic	1072253.75	93952
84837	20175355001	Produce	Fruit	Bananas	Yellow	1052886.84	737799
61325	21087193	Meat	Fresh-Poultry	Fresh-Poultry	Breasts	775428.01	68869
92721	20425775001	Produce	Fruit	Grapes	Green Grapes	756235.44	119984
8940	20252014	HMR	HMR	Ready to Eat	Whole Rotisserie Chicken	755620.45	63705
15155	20600985	Deli	Gourmet Foods	Gourmet Foods	Olives	673152.02	93454
23863	20794110	Meat	Fresh Beef	Fresh-Beef	AAA	641822.43	9145
84569	20159199001	Produce	Fruit	Grapes	Red Grapes	640238.30	113183
25710	20821992	Meat	Fresh-Poultry	Fresh-Poultry	Breasts	619704.29	35152
8557	20188873	Dairy	Milk & Eggs	Milk	Core Milk	612231.33	139919

b.



- The process started by calculating total revenue and transaction volume for each product across the entire dataset. The data were then aggregated, merged to juxtapose revenue and volume, and sorted to determine the top performers.

- The analysis uncovered that products like electronics and certain produce items, despite being less frequently sold than others, contribute significantly to total revenue. This demonstrates the critical role of high-value and high-margin items in ACSE's sales strategy.

- The evidence for this insight is derived from the top 10 products by revenue and their corresponding transaction volumes, as displayed in the bar chart. The contrast between revenue and volume visually underscores the impact of these high-value products.

- Regarding profitability, the evidence suggests that while some items may have fewer transactions, their high price points result in substantial revenue, which is a common characteristic of profitable product strategies. It underscores the importance of a balanced inventory that includes both high-turnover items for steady cash flow and high-margin items for substantial profit contributions, reflecting a well-rounded approach to ACSE's inventory management strategy.

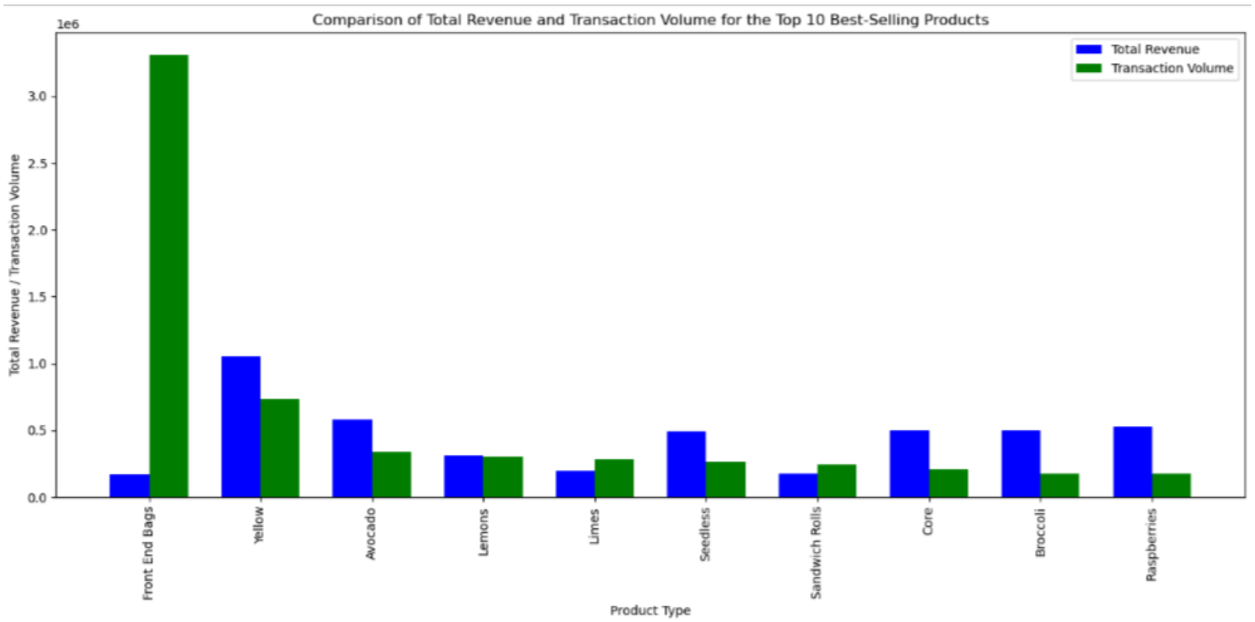
  ○ Show the product distribution by transaction volume.

a.

```
Top 10 products by transaction volume:
```

	prod_id	prod_section	prod_category	prod_subcategory	prod_type	total_revenue	transaction_volume
8570	20189092	Home	Household	Front End Bags	Front End Bags	165750.63	3309848
84837	20175355001	Produce	Fruit	Bananas	Yellow	1052886.84	737799
101574	21097012001	Produce	Fruit	Tropical	Avocado	579378.83	336738
82645	20028593001	Produce	Fruit	Citrus	Lemons	313774.12	305189
82842	20040489001	Produce	Fruit	Citrus	Limes	193373.91	281332
83242	20070132001	Produce	Vegetables	Field Veg	Seedless	493495.63	260311
3469	20076950	Bakery Instore	In-Store	Rolls-In-Store	Sandwich Rolls	176265.85	245265
97272	20812144001	Dairy	Milk & Eggs	Eggs	Core	498331.47	210452
84346	20145621001	Produce	Vegetables	Cooking Veg	Broccoli	501997.65	176687
84116	20128938001	Produce	Fruit	Berries/Cherries	Raspberries	525740.69	174815

b.



- This involved plotting a side-by-side bar chart to compare total revenue and transaction volume for ACSE's top 10 best-selling products, highlighting differences in sales performance metrics.

- The visualization indicates that products like 'Front End Bags' lead in transaction volume but are not the highest revenue generators, suggesting the prominence of certain items in quantity sold over the revenue contribution.

- The provided bar chart serves as evidence for this insight, clearly demonstrating how some products generate substantial transaction volume, which can be strategic for attracting customers to the store, even if they don't contribute the most to revenue.

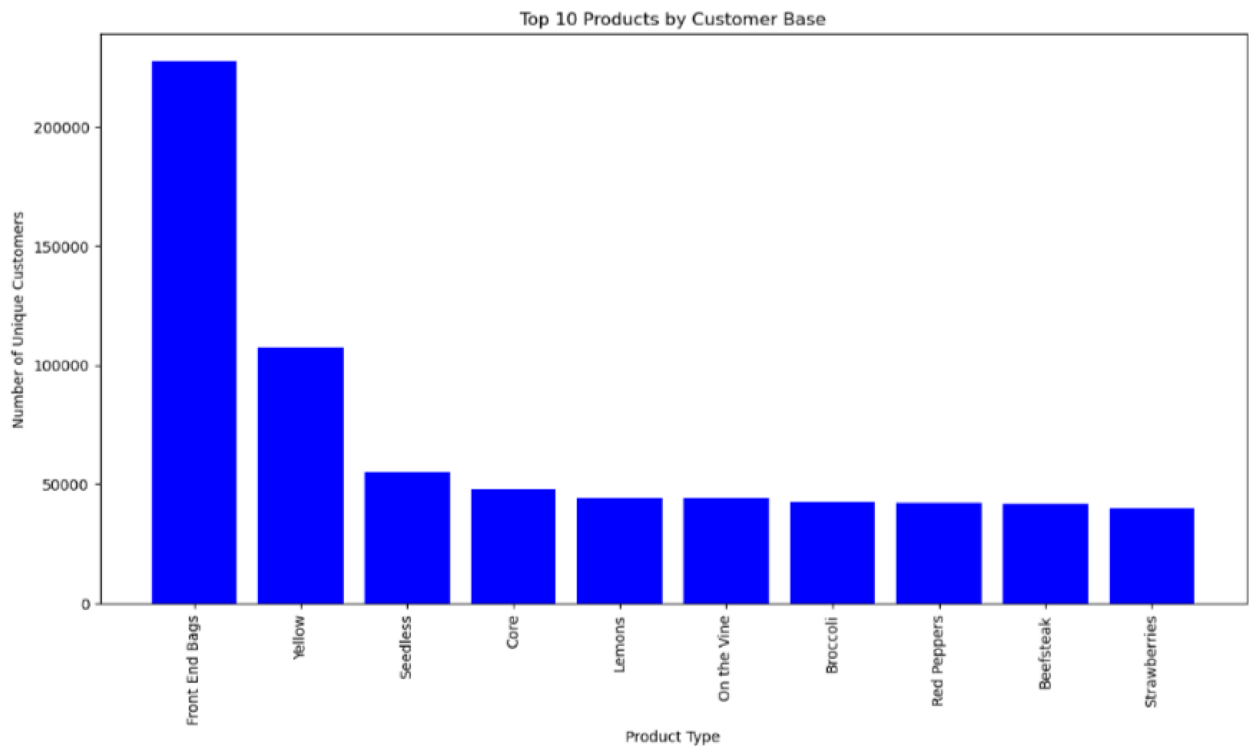
- This finding can be significant for ACSE's business strategy. High-transaction-volume products might offer lower revenue per unit but could be key drivers for store traffic, which offers opportunities for cross-selling and up-selling higher-margin items. These products might also serve as 'loss leaders', which are sold at a low price to attract customers, who may then purchase additional, more profitable items. Maintaining a strategic mix of these products can optimize overall profitability through increased customer footfall and the potential for additional sales

4. The products with the most customer base.

a.

	prod_section	prod_category	prod_subcategory	prod_type	prod_id	unique_customers	customer_percentage
62349	Home	Household	Front End Bags	Front End Bags	20189092	227743	56.280087
98468	Produce	Fruit	Bananas	Yellow	20175355001	107234	26.499778
99558	Produce	Vegetables	Field Veg	Seedless	20070132001	54899	13.566698
7127	Dairy	Milk & Eggs	Eggs	Core	20812144001	48174	11.904809
98550	Produce	Fruit	Citrus	Lemons	20028593001	44340	10.957347
100223	Produce	Vegetables	Tomatoes	On the Vine	20026703001	44287	10.944249
99261	Produce	Vegetables	Cooking Veg	Broccoli	20145621001	42428	10.484851
99852	Produce	Vegetables	Peppers	Red Peppers	20007535001	42064	10.394899
100197	Produce	Vegetables	Tomatoes	Beefsteak	20426141001	41700	10.304947
98522	Produce	Fruit	Berries/Cherries	Strawberries	20049778001	40123	9.915237

b.



The strategy involved first determining the total number of unique customers and then grouping the data by product hierarchy to count the unique customers for each product. This was followed by calculating the percentage of the total customer base that each product attracted.

The analysis shows that 'Front End Bags' have the largest customer base, with **56.28%** of ACSE's total customers purchasing them, indicating that a significant portion of customers buy these items.
This finding is supported by the bar chart and sorted data, where 'Front End Bags' lead in the number of unique customers, and the calculation that more than half of ACSE's customer base has purchased them.

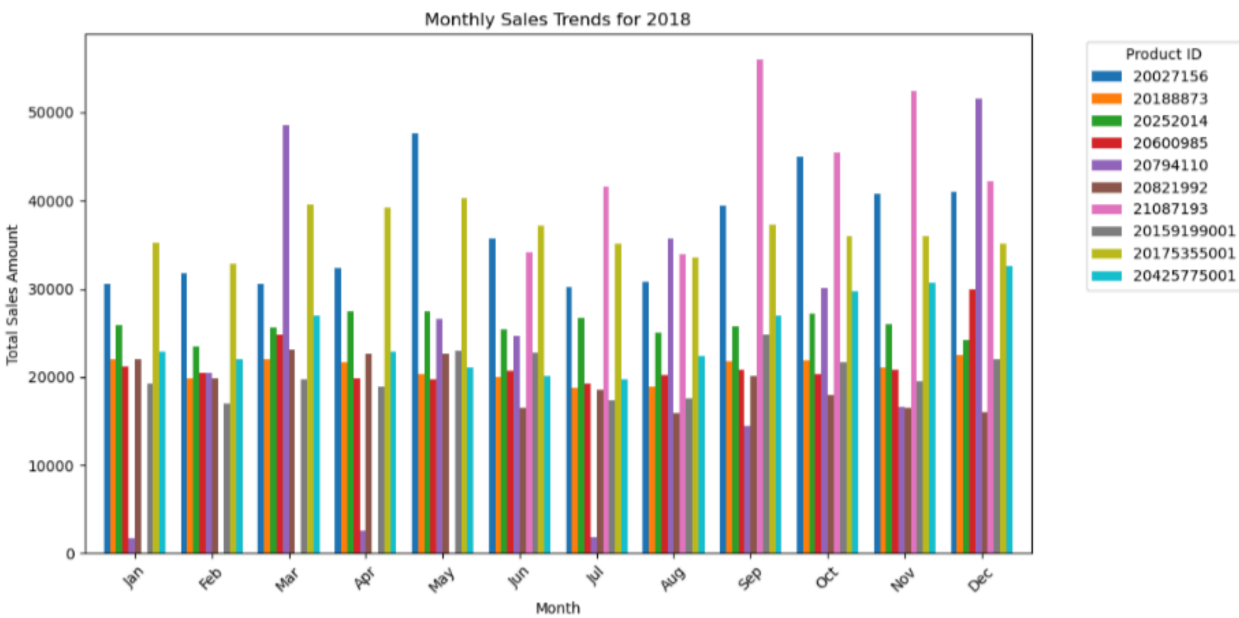
The substantial customer base for 'Front End Bags' suggests they are a common and possibly essential product for a majority of ACSE's shoppers. This could imply that while some products may not generate the highest revenue per unit, their role in attracting customers to the store is critical, and they may serve as a gateway to the purchase of additional items, enhancing overall sales and providing opportunities for cross-promotion and upselling.

5. Monthly sales trends of the two years

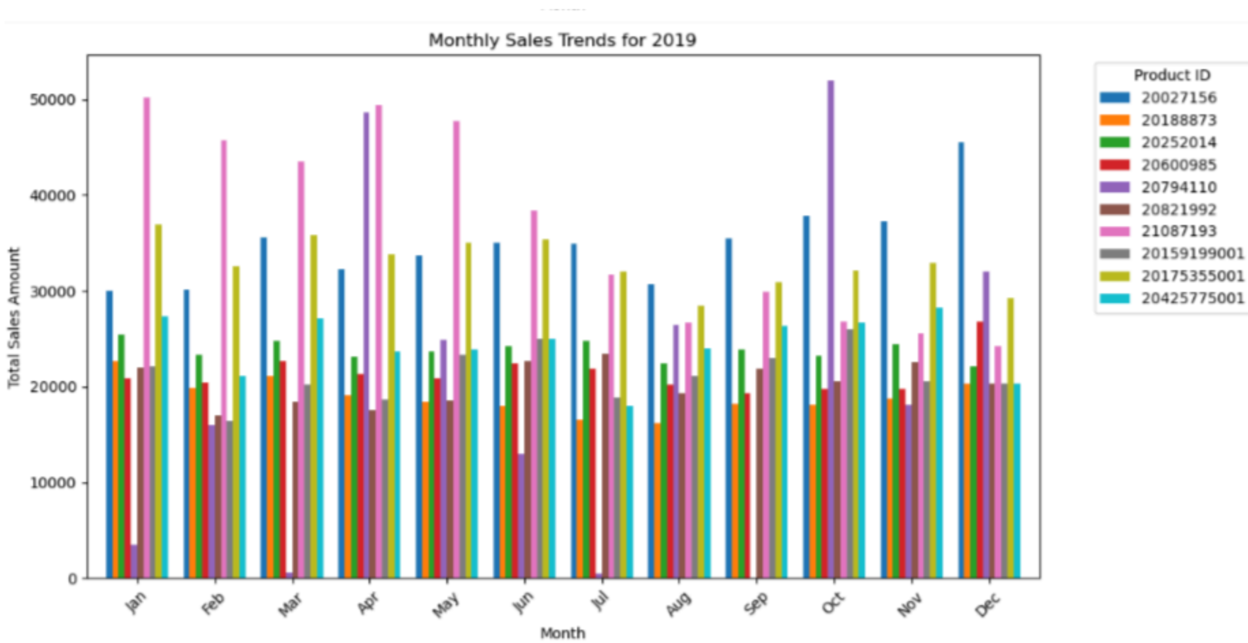
a.

	prod_id	prod_section	prod_category	prod_subcategory	prod_type	total_revenue
1218	20027156	Customer Service	Lottery - Electronic	LOTTERY - ELECTRONIC	Electronic	1072253.75
84837	20175355001	Produce	Fruit	Bananas	Yellow	1052886.84
61325	21087193	Meat	Fresh-Poultry	Fresh-Poultry	Breasts	775428.01
92721	20425775001	Produce	Fruit	Grapes	Green Grapes	756235.44
8940	20252014	HMR	HMR	Ready to Eat	Whole Rotisserie Chicken	755620.45
15155	20600985	Deli	Gourmet Foods	Gourmet Foods	Olives	673152.02
23863	20794110	Meat	Fresh Beef	Fresh-Beef	AAA	641822.43
84569	20159199001	Produce	Fruit	Grapes	Red Grapes	640238.30
25710	20821992	Meat	Fresh-Poultry	Fresh-Poultry	Breasts	619704.29
8557	20188873	Dairy	Milk & Eggs	Milk	Core Milk	612231.33

b.



c.



- The approach was to extract transaction data for each product by month for the years 2018 and 2019, ensuring that only complete years with data for all months were analyzed to maintain consistency and accuracy.

- Specific products such as Electronic Lottery Tickets (**ID 20027156**) and Bananas (**ID 20175355001**) show marked sales peaks during certain months, indicating a strong seasonal or event-driven demand. On the other hand, products like Front End Bags (**ID 20189901**) and Whole Rotisserie Chicken (**ID 20252014**) demonstrate consistent sales, suggesting they are staple goods with a steady customer base.

- The sales peaks for Electronic Lottery Tickets in October of both years could point to a pattern, perhaps a recurring event or a holiday-specific promotion. Meanwhile, the constant demand for Front End Bags across all months highlights their essential role in shopping activities. The summer sales increase for Bananas and the consistent performance of Whole Rotisserie Chicken further illustrate seasonal buying trends and everyday consumer habits.

- These sales patterns offer strategic insights for ACSE. For instance, the seasonal peaks in sales for Electronic Lottery Tickets could inform targeted marketing during those high-demand periods, potentially increasing profitability through promotional activities. The steady sales of staples like Whole Rotisserie Chicken and Front End Bags underscore the importance of maintaining adequate stock levels to meet the consistent demand, thus ensuring a continuous revenue flow. Understanding these trends enables ACSE to optimize inventory, manage supply chain logistics, and tailor marketing strategies to enhance customer satisfaction and profitability.

6. **Products that have quality-issue (aka. High return rate; negative sales_amt or sales_wgt)**

a.

```
Number of transactions with negative sales_amt: 651001
Number of transactions with negative sales_wgt: 3286
```

- A large number of negative sales_amt transactions strongly suggest a return process is captured in the data. Customers are likely returning products, which results in negative revenue for those transactions. Whereas the lower number of negative sales_wgt transactions might indicate that not all products have weights associated with them or that weights are not always recorded when products are returned.

b.

	prod_section	prod_category	prod_subcategory	prod_type	prod_id	total_sales_amt	total_returned_amt	total_sales_qty	total_returned_qty	return_rate_by_amt	return_rate_by_qty
339	Baby	Baby	Baby Accessories	Potty	20183200001	16.99	-16.99	-1	-1	1.000000	1.000000
675	Baby	Baby	Baby Toiletries	Toddler	21194299	5.99	-5.99	-1	-1	1.000000	1.000000
7412	Dairy	Milk & Eggs	Milk	Premium Milk	20057494	5.89	-5.89	-1	-1	1.000000	1.000000
11951	Entertainment	Photo Image	Off-Site	Off-site	20784140	60.00	-60.00	-1	-1	1.000000	1.000000
12595	Entertainment	Reading	Books-Adult	Non Fiction	20965832	11.39	-11.39	-1	-1	1.000000	1.000000
...	...	...	...	...	...	...	...	...	...	...	...
76064	Mass Cosmetics	Colour Cosmetics - Mass	Eye Colour	Eye Shadow	20733363002	9.79	-5.90	1	0	0.602656	0.000000
72118	Home	Soft Goods (Textiles)	Sheets	Duvet Covers	21006673	124.98	-74.99	0	-1	0.600016	0.000000
62259	Home	Household	Foil-Household	Baking	20941884	9.26	-5.50	3	-1	0.593952	0.333333
67441	Home	Kitchen Prep	Serveware	Platters And Trays	20939549	321.34	-182.04	1	-12	0.566503	12.000000
49584	HBA	Grooming	Shaving Products	Hair color	20986065	23.98	-12.99	0	-1	0.541701	0.000000

100 rows × 11 columns

- This table reveals that certain products, notably '**Baby Accessories Bath 21067490**', '**Baby Accessories Bath 21067549**', and 'Baby Accessories Bath 21067215', exhibit exceptionally high return rates. This trend indicates potential quality or customer satisfaction issues, suggesting that these products should be prioritized for a comprehensive quality review and corrective action to mitigate the high incidence of returns.

- The method included identifying negative values in the 'sales_amt' and 'sales_wgt' fields, presuming these indicate product returns, and then calculating the return rate by both amount and quantity to assess the extent of returns for each product.

- The analysis indicated a substantial number of transactions with negative 'sales_amt', suggesting a significant return rate that could point to quality issues or customer dissatisfaction with certain products. However, a much lower number of transactions with negative 'sales_wgt' implies that weight may not always be recorded during the return process.

- The evidence is provided by the data, showing **651,001** transactions with negative sales amounts and only **3,286** with negative weights. This discrepancy suggests that while returns are common, they may not always include the product weight, potentially due to the nature of the products or the return process itself.

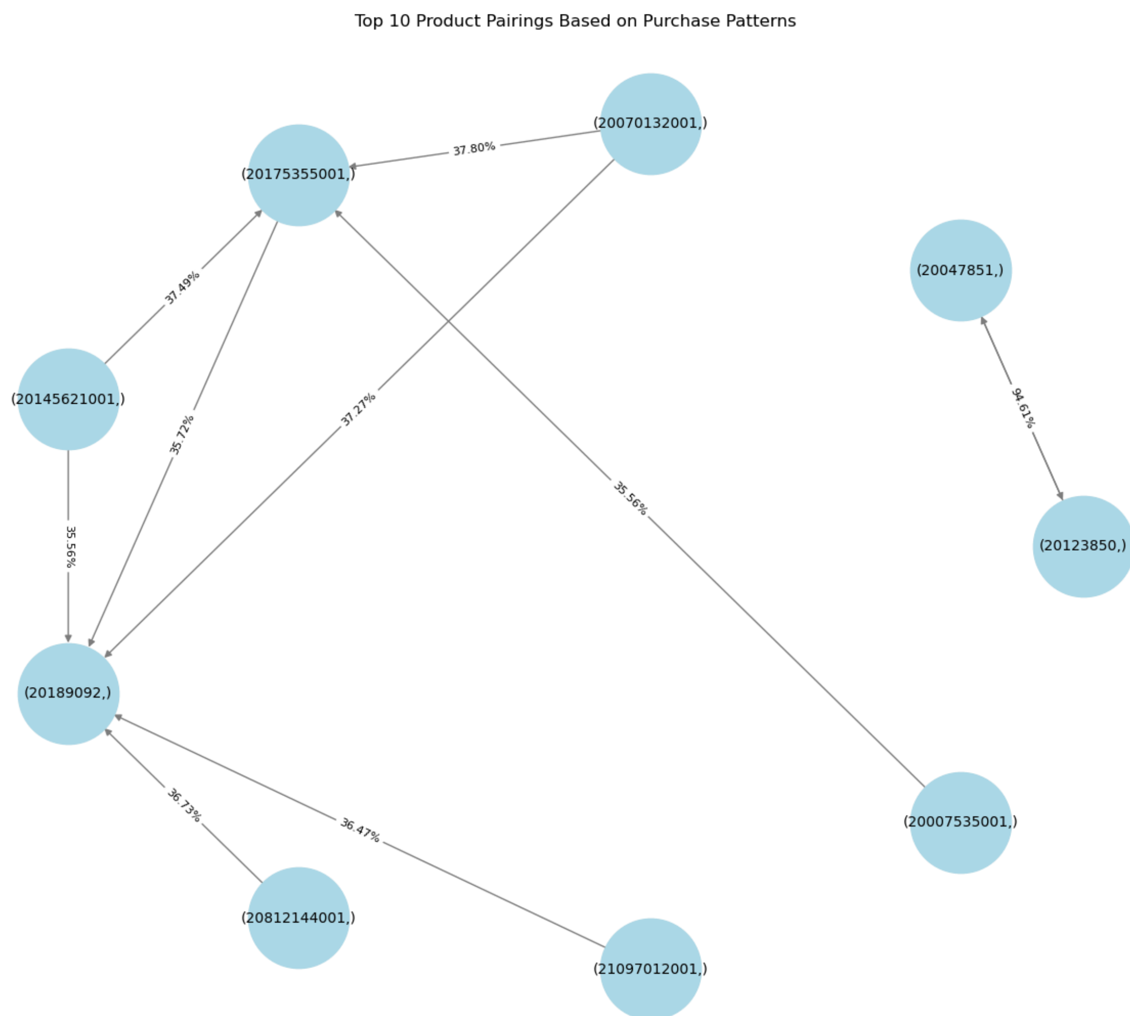
- Products with high return rates, as reflected by the total return amounts and quantities, may represent a quality or satisfaction issue for ACSE. Identifying these products is crucial for addressing potential quality control problems, improving customer satisfaction, and ultimately reducing return rates, which can positively impact profitability. Lower return rates often correlate with higher customer retention and lower costs associated with processing returns, which in turn can improve the company's financial health.

7. The products that have the highest cross-selling rates

a.

```
{20123850} -> {20047851} (conf: 0.995, supp: 0.022, lift: 42.215, conv: 178.460)
{20047851} -> {20123850} (conf: 0.946, supp: 0.022, lift: 42.215, conv: 18.126)
{20070132001} -> {20175355001} (conf: 0.378, supp: 0.015, lift: 2.890, conv: 1.397)
{20145621001} -> {20175355001} (conf: 0.375, supp: 0.011, lift: 2.866, conv: 1.391)
{20070132001} -> {20189092} (conf: 0.373, supp: 0.014, lift: 1.214, conv: 1.105)
{20812144001} -> {20189092} (conf: 0.367, supp: 0.012, lift: 1.196, conv: 1.095)
{21097012001} -> {20189092} (conf: 0.365, supp: 0.010, lift: 1.188, conv: 1.091)
{20175355001} -> {20189092} (conf: 0.357, supp: 0.047, lift: 1.164, conv: 1.078)
{20145621001} -> {20189092} (conf: 0.356, supp: 0.010, lift: 1.158, conv: 1.075)
{20007535001} -> {20175355001} (conf: 0.356, supp: 0.010, lift: 2.719, conv: 1.349)
```

b.



- The analysis used the apriori algorithm on transactional data to identify products commonly purchased together. This method assessed the likelihood of certain products being bought in tandem by evaluating the strength of association between item pairs.

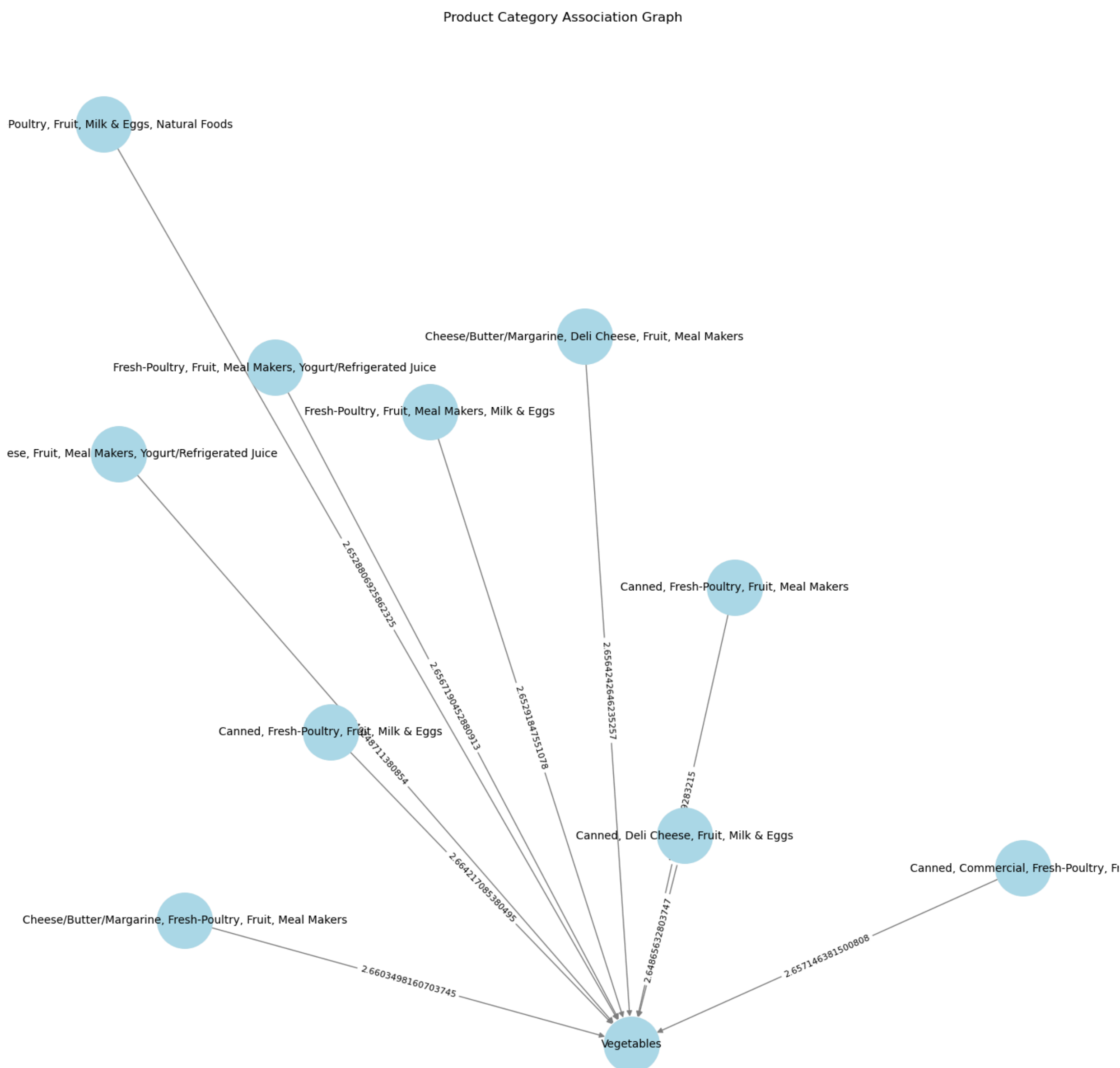
- Insight: The analysis determined that certain product pairs, such as those including IDs {**20123850, 20047851**}, have exceptionally high cross-selling rates, with a confidence level as high as 0.995, suggesting customers who purchase one item are very likely to also purchase the other.

- The evidence comes from the apriori output, showing strong association rules with high confidence values. For example, when customers buy the product with **ID 20123850**, there is a **99.5%** chance they will also buy product **ID 20047851**, which is indicated by a confidence level of 0.995 and a lift value far exceeding 1 (**42.215**), signifying a strong positive relationship between these items.

- The visual representation in the network graph confirms these associations, with directed edges representing the direction of the association from the antecedent to the consequent product. The width of the edges reflects the confidence level, portraying the strength of the association. Products that show such strong connections to others in customer purchases are ideal candidates for bundling strategies, targeted promotions, and cross-selling efforts, all of which can potentially increase the overall basket size and revenue. Identifying high cross-selling rates allows ACSE to better understand customer purchasing patterns and can influence stocking and sales strategies to maximize profitability.

8. Correlation between product categories.

a.



- Utilizing the apriori algorithm, the analysis was performed on transactional data to discover correlations between product categories based on their co-occurrence in the same transactions.

- The network graph generated from the apriori analysis indicates strong associations between certain product categories. For instance, **'Fresh Poultry**', **'Fruit', 'Meal Makers**', **'Milk & Eggs', and 'Natural Foods**' categories are frequently purchased together, suggesting a common shopping pattern among customers.

- The evidence is visually represented in the Product Category Association Graph, which illustrates the strength of associations between product categories through directed edges. Categories such as '**Canned**', '**Deli Cheese**', and '**Fresh Poultry**' show a strong correlation, as indicated by the proximity and interconnectedness in the network graph.

- The graph suggests that certain categories consistently appear together in customer baskets, indicating possible meal planning or common culinary uses. These insights could be valuable for strategic product placement, bundle promotions, and inventory management. For example, placing items from complementary categories in proximity or running promotions that span these categories could increase basket size and enhance customer experience. Understanding these correlations can help ACSE tailor its marketing strategies to capitalize on these shopping patterns, potentially increasing sales and customer satisfaction.

9. What % of transactions are ACSE-made products?

a.

```
Percentage of transactions with ACSE-made products: 79.80%
```

- The method involved filtering the dataset for transactions that included ACSE-made products, counting unique transactions containing these products, and then calculating this as a percentage of all transactions.

- The analysis revealed that ACSE-made products constitute 79.8% of all transactions, indicating a substantial brand presence within the sales data.

- The evidence for this substantial brand penetration is the calculated percentage of transactions involving ACSE products, highlighting the brand's dominant role in the product mix offered to customers.

- This high percentage of transactions with ACSE-made products underscores the brand's strong market presence and suggests that the majority of customers purchase these products. This could be due to various factors, such as consumer loyalty, competitive pricing, product quality, or effective branding and marketing strategies. The prominence of ACSE-made products in sales transactions can be a significant advantage in market positioning and provides opportunities to leverage the brand further to enhance customer retention and attract new customers.

3. <u>Store</u>

Based on the analysis of the 58 stores, our team's investigation into store performance revealed a positive correlation between the revenue of a store and its transaction volume. The top and bottom 20 stores, by both transaction volume and revenue, were identified and cross-referenced through store_id. This analysis showed that the majority of stores have similar rankings in both metrics, underscoring the crucial role of transaction volume in driving a store's revenue.

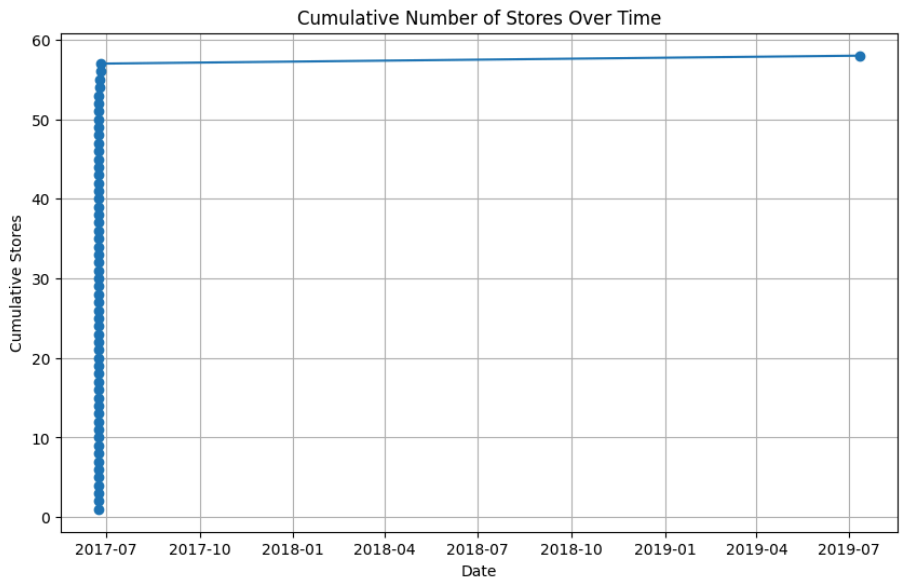
```
Top stores by revenue:
+--------+--------------------+
|store_id|       total_revenue|
+--------+--------------------+
|    1212|1.0651950299998906E7|
|    1050|   9526095.269999478|
|    1007|   8788538.759999854|
|    1004|    8544597.63999998|
|    1066|   8430151.460000057|
|    1021|   8327886.240000117|
|    1035|   7947683.870000289|
|    1027|   7308581.490000578|
|    1188|   7047047.000000761|
|    1011|   6982799.220000594|
|    1040|   6787875.020000796|
|    1051|   6411543.120000845|
|    1114|   6361870.640000747|
|    1016|    6270114.49000075|
|    1019|   6199663.010000778|
|    1029|   6060483.960000788|
|    1028|   6009527.590000684|
|    1014|   5885459.650000732|
|    1001|   5841673.250000653|
|    1079|  5751749.3800006695|
+--------+--------------------+
only showing top 20 rows
```

```
Bottom stores by revenue:
+--------+------------------+
|store_id|     total_revenue|
+--------+------------------+
|    5264|1409.4099999999999|
|    1223|           1720.67|
|    1214|            2654.9|
|    1220|2674.1799999999994|
|    1211|           3690.05|
|    1227|           3986.34|
|    1217|           4174.44|
|    1222|           5067.45|
|    1210| 5694.360000000001|
|    1231| 5742.970000000001|
|    1213|5826.1100000000015|
|    1221|           6958.78|
|    1142|1130024.1900000297|
|    1179| 1728043.050000091|
|    1200|1828468.8400000774|
|    1174|2312897.7600000273|
|    1023| 2404439.190000016|
|    1154| 2431191.860000013|
|    1208|2452100.5100000193|
|    1132| 2653452.289999989|
+--------+------------------+
only showing top 20 rows
```

```
Bottom stores by transaction volume:
+--------+------------------+
|store_id|transaction_volume|
+--------+------------------+
|    1231|               683|
|    1223|               787|
|    1214|              1004|
|    1213|              1010|
|    1211|              1240|
|    1222|              1573|
|    1227|              1711|
|    1220|              1743|
|    1217|              2063|
|    1210|              2067|
|    1221|              2108|
|    5264|              2479|
|    1142|            257921|
|    1179|            374802|
|    1200|            381607|
|    1208|            434757|
|    1023|            459568|
|    1174|            475611|
|    1154|            494558|
|    1132|            558448|
+--------+------------------+
only showing top 20 rows
```

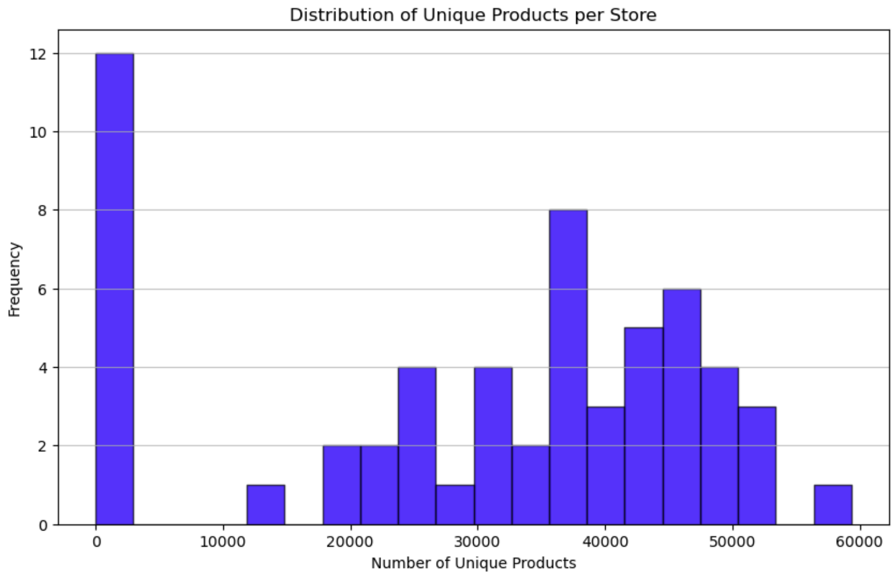
```
Top stores by transaction volume:
+--------+------------------+
|store_id|transaction_volume|
+--------+------------------+
|    1212|           1962183|
|    1007|           1758017|
|    1050|           1755499|
|    1004|           1555537|
|    1066|           1509869|
|    1021|           1455833|
|    1035|           1404113|
|    1027|           1363141|
|    1040|           1285805|
|    1188|           1282754|
|    1011|           1200371|
|    1051|           1172923|
|    1019|           1171206|
|    1016|           1166595|
|    1114|           1160444|
|    1010|           1113772|
|    1029|           1113192|
|    1000|           1087753|
|    1014|           1080052|
|    1079|           1056372|
+--------+------------------+
only showing top 20 rows
```

An examination of the store opening patterns from 2017 to 2022 indicated a very gradual expansion, with only one new store opening, suggesting recent stagnation in growth. This points to potential underlying issues that could be hindering the brand's development.



Our team's analysis of product diversity across stores has concluded that the most profitable stores do not necessarily carry a wider range of unique products compared to the average. This suggests that there may be a point of diminishing returns when it comes to the number of unique products offered. Essentially, expanding product lines may not always correlate with increased revenue and could potentially add to management complexity and costs. The distribution of product quantities across stores, illustrated by histograms, could inform management decisions to optimize inventory levels, ensuring the availability of high-demand products and reducing stock of less popular items. Even if certain stores offer fewer unique products than the average, it does not hinder their ability to generate high revenue, implying that these stores may focus on selling high-margin items or that their product mix is more aligned with their target market's preferences.

```
Average number of unique products per store: 30106.91379310345
```



Iv.

In assessing store efficiency, we examined revenue generation in relation to store size. Interestingly, we found no direct correlation between store size and revenue, implying that larger stores do not necessarily report higher earnings. This aligns with the conventional wisdom that a multitude of factors, including store location and customer demographics, play a more significant role than size alone. Therefore, it's crucial to analyze additional variables such as store location, consumer traffic, and local competition to understand the dynamics of revenue per square foot fully.

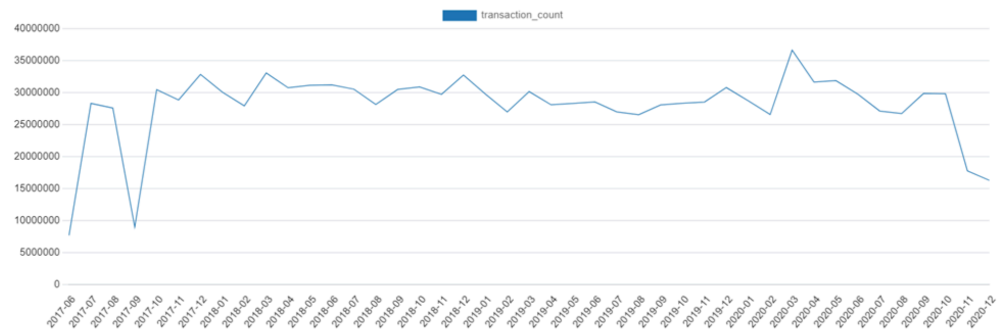
	store_id	unique_products	total_revenue	estimated_store_size	revenue_per_sq_ft
47	1212	46963	10651950.30	469630	22.681580
5	1007	43017	8788538.76	430170	20.430385
3	1004	44365	8544597.64	443650	19.259772
42	1194	29332	5555563.02	293320	18.940280
57	5264	8	1409.41	80	17.617625
24	1066	47885	8430151.46	478850	17.604994
30	1095	23340	4057392.26	233400	17.383857
12	1021	48051	8327886.24	480510	17.331348
19	1035	46877	7947683.87	468770	16.954336
0	1000	30758	5211321.95	307580	16.942981
32	1114	38436	6361870.64	384360	16.551854
17	1029	37036	6060483.96	370360	16.363765
21	1050	59339	9526095.27	593390	16.053684
15	1027	46563	7308581.49	465630	15.696114
41	1188	46298	7047047.00	462980	15.221061
10	1019	42480	6199663.01	424800	14.594310
25	1079	39813	5751749.38	398130	14.446913
20	1040	46994	6787875.02	469940	14.444131
6	1010	38415	5504877.09	384150	14.330020
31	1099	36241	5137445.86	362410	14.175784
37	1155	26451	3653434.04	264510	13.812083
11	1020	21528	2908651.56	215280	13.511016
8	1014	44382	5885459.65	443820	13.260916
7	1011	52857	6982799.22	528570	13.210737
29	1092	35850	4622972.40	358500	12.895321
4	1005	39847	5130519.11	398470	12.875547
1	1001	45509	5841673.25	455090	12.836303
22	1051	50104	6411543.12	501040	12.796470
28	1090	33421	4212597.64	334210	12.604643

33	1127	36576	4540926.52	365760
38	1170	33217	4089613.68	332170
9	1016	51125	6270114.49	511250
13	1022	38017	4609335.75	380170
39	1174	19529	2312897.76	195290
16	1028	51265	6009527.59	512650
27	1083	43968	5072676.00	439680
23	1064	31597	3580972.85	315970
18	1032	48335	5351867.71	483350
34	1132	24108	2653452.29	241080
36	1154	24872	2431191.86	248720
2	1003	41168	3774978.03	411680
35	1142	12763	1130024.19	127630
40	1179	19530	1728043.05	195300
26	1082	36739	3061823.95	367390
14	1023	30404	2404439.19	304040
43	1200	23934	1828468.84	239340
44	1208	32634	2452100.51	326340
56	1231	348	5742.97	3480
48	1213	375	5826.11	3750
52	1221	498	6958.78	4980
53	1222	436	5067.45	4360
46	1211	330	3690.05	3300
45	1210	510	5694.36	5100
55	1227	409	3986.34	4090
50	1217	436	4174.44	4360
49	1214	326	2654.90	3260
51	1220	336	2674.18	3360
54	1223	256	1720.67	2560

Data Cleaning and Sampling Method

Due to the substantial size of the transactions table, comprising 120 million rows, executing even basic queries demanded considerable computational resources. To make data processing manageable, we created a representative sample dataset for the transactions table. This approach gave us profound insights while reducing the overall data size for more efficient analysis.

1. **Run simple queries on entire tables using Postgre SQL**
   First, we conducted Exploratory Data Analysis (EDA) on the entire tables to gain a rough understanding of the data, including data size, general trends over time, and product category hierarchy.
   - Transactions table
     - Finding 1: Transaction volume remained stable over the years until a significant shift occurred in February 2020 due to COVID-19.



     - Finding 2: There is 1 store, #8540, that has its last transaction date in 2019. We assume this store is closed.

store_id	max_trans_dt
8540	**6/9/2019**
1213	**8/22/2020**
1210	**11/22/2020**
1211	**11/22/2020**
1220	**11/22/2020**
1217	12/21/2020
1000	12/24/2020
1001	12/24/2020
1004	12/24/2020
1005	12/24/2020

   - Products table
     - Finding 1: Non-products, defined as items that are not the main purpose of customers' store visit, are included in the products table (e.g. plastic bags, coupons)

- ○ Finding 2: The product category consists of multiple layers: *section > category > subcategory > type > product*

	unique_products bigint	unique_sections bigint	unique_categories bigint	unique_subcategories bigint	unique_types bigint
1	154818	33	101	430	2010

2. **Reduce the size of the Transactions table and download it onto our local computer**
We weren't allowed to create a new table or delete rows in the Postgre database. Given technical constraints, we decided to store data outside the Postgres database, manipulate the data with Python, and create a sample. Due to the significant size of the transactions table, downloading the data onto our local computers was very time-consuming. Therefore, leveraging the simple insights gained from step 1, we trim down the dataset size and download the transactions table on our computer.

- Steps
  - ○ Remove transactions from February 2020 onward (covid era) since they have unusual trends.
  - ○ Remove all transactions from store 8540 since it's been closed.

3. **Make transactions sample**
We sampled transaction data by randomly selecting customers and retrieving the complete transaction history of those chosen customers. To construct an accurate recommendation system, it's crucial to preserve the customer's entire purchasing history. When selecting customers randomly, we assigned an index to each unique customer_id and made selections based on these indices. This approach was necessary because customer_id is not randomly generated, as previously mentioned, leading to the hashing algorithm selecting customers non-randomly.

- Steps
  - ○ Assign an index to each unique customer
  - ○ Select 5% of unique customers based on their index
  - ○ Get all the transaction history of selected customers.

4. **Drop duplicates from a sample**
There are duplicates in the sample data, meaning that there are rows with the same trans_id and prod_id. We dropped those duplicates from the sample.

5. **Join the Products and Transactions table for EDA.**
After having a complete sample transaction table, we joined it with the products table for exploratory data analysis.

Future Work

Through analysis, we were able to fully understand critical insights of ACSE’s customers, products, and stores. Our findings will serve as the foundation for the development of a recommender system tailored to ACSE’s needs. Looking ahead, our work will focus on two key areas to further refine and implement the recommender system:

(1) Text Mining for Refined Categorization: We will employ advanced text mining techniques to analyze product descriptions. This will enable us to uncover additional product groupings beyond the existing categories, potentially revealing hidden patterns in product groups and customer preferences. This refined categorization will enhance the personalization capabilities of the recommender system, allowing for more targeted and relevant product recommendations.

(2) Development and Implementation of the Recommender System: Informed by insights such as top-selling products, current purchasing behaviors, and common purchase patterns within customer groups, we will initiate the development of a recommender system tailored to ACSE's strategic needs. This system will utilize sophisticated algorithms to forecast and recommend products, optimizing decisions related to inventory selection, shelf space allocation, and promotional activities. By aligning product offerings with individual customer preferences, the system aims to refine the shopping experience and enhance operational efficiencies,.