



E-commerce Analysis

By NC

SQL

Big Query

E-commerce SQL Analysis

Context:

The e-commerce industry is highly dynamic, with customer preferences and market trends constantly evolving. To stay ahead, companies must continuously analyse their sales, product, and customer data to identify key performance indicators (KPIs) and patterns that can influence strategic decisions. This project focuses on utilizing SQL to explore and analyse various datasets—demographic, transaction, and product data—to uncover actionable insights. These insights will be pivotal in understanding customer behaviour, optimizing product offerings, and enhancing overall business performance.

Objective:

The primary objective of this analysis is to harness the power of SQL to explore e-commerce data and uncover patterns, trends, and key performance indicators that contribute to profitable growth. Specific goals include:

- Analyzing customer demographics to identify segments that drive the most value.
- Investigating sales data to determine revenue drivers and optimize product offerings.
- Evaluating product performance to understand trends and customer preferences.
- Calculating KPIs such as average basket size, customer retention rates, and sales per product category.

Approach:

Using SQL in Big Query, the analysis will involve querying the provided datasets to extract relevant metrics, perform aggregations, and generate detailed reports. This approach will enable a comprehensive understanding of the data, facilitating the identification of opportunities for improving customer engagement, optimizing product strategies, and driving overall business growth.

Data Dictionary:

Demographic table

Variable	Description
HOUSEHOLD_KEY	Uniquely identifies each household
AGE_DESC	Estimated age range
MARITAL_STATUS_CODE	Marital Status (A - Married, B- Single, U - Unknown)
INCOME_DESC	Household income
HOMEOWNER_DESC	Homeowner, renter, etc.
HH_COMP_DESC	Household composition
HOUSEHOLD_SIZE_DESC	Size of household up to 5+
KID_CATEGORY_DESC	Number of children present up to 3+

Transaction table:

Variable	Description
HOUSEHOLD_KEY	Uniquely identifies each household
BASKET_ID	Uniquely identifies a purchase occasion
DAY	Day when transaction occurred
PRODUCT_ID	Uniquely identifies each product
QUANTITY	Number of the products purchased during the trip
SALES_VALUE	Amount of dollars retailer receives from sale
STORE_ID	Identifies unique stores
COUPON_MATCH_DISC	Discount applied due to retailer's match of manufacturer coupon
COUPON_DISC	Discount applied due to manufacturer coupon
RETAIL_DISC	Discount applied due to retailer's loyalty card program
TRANS_TIME	Time of day when the transaction occurred
WEEK_NO	Week of the transaction. Ranges 1 - 102

Products Table:

Variable	Description
PRODUCT_ID	Number that uniquely identifies each product
DEPARTMENT	Groups similar products together
COMMODITY_DESC	Groups similar products together at a lower level
SUB_COMMODITY_DESC	Groups similar products together at the lowest level
MANUFACTURER	Code that links products with same manufacturer together
BRAND	Indicates Private or National label brand
CURR_SIZE_OF_PRODUCT	Indicates package size (not available for all products)

KPIs & Metrics

1. Total Sales

```
select cast(round(sum(SALES_VALUE)) as integer) as Total_Sales
from `ecommerce.transaction_data`
```

Row	Total_Sales
1	4029338

2. Total Number of Orders

```
select count(distinct BASKET_ID) as Total_Orders
from `ecommerce.transaction_data`
```

Row	Total_Orders
1	233356

3. Average Order Value

```
select cast(round(sum(SALES_VALUE)/count(distinct BASKET_ID)) as
integer) as Avg_Order_Value
from `ecommerce.transaction_data`
```

Row	Avg_Order_Value
1	17

4. Total Number of Products

```
select count(distinct SUB_COMMODITY_DESC) as Total_Products
from `ecommerce.product`
```

Row	Total_Products
1	2383

5. Total Number of Households

```
select count(distinct household_key) as Total_Households
from `ecommerce.hh_demographic`
```

Row	Total_Households
1	801

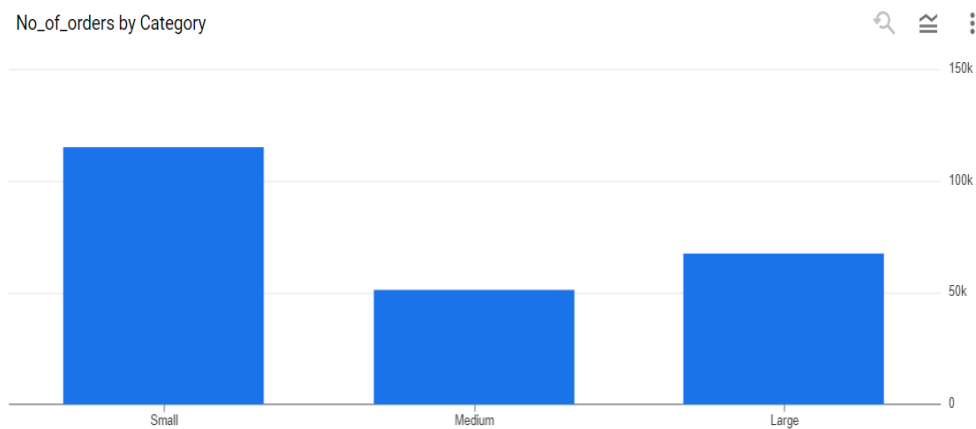
Queries

Question 1: Find the number of orders that have small, medium or large order value (small:0-10 dollars, medium:10-20 dollars, large:20+)

```
WITH cte AS
(SELECT BASKET_ID, SUM(SALES_VALUE) AS order_value
FROM `ecommerce.transaction_data`
GROUP BY BASKET_ID)
SELECT 'Small' AS Category, COUNT(*) AS No_of_orders
FROM cte
WHERE order_value < 10
UNION ALL
SELECT 'Medium' AS Category, COUNT(*) AS No_of_orders
FROM cte
WHERE order_value >= 10 AND order_value <= 20
UNION ALL
SELECT 'Large' AS Category, COUNT(*) AS No_of_orders
FROM cte
WHERE order_value > 20;
```

Row	Category	No_of_orders
1	Small	115045
2	Medium	51000
3	Large	67311

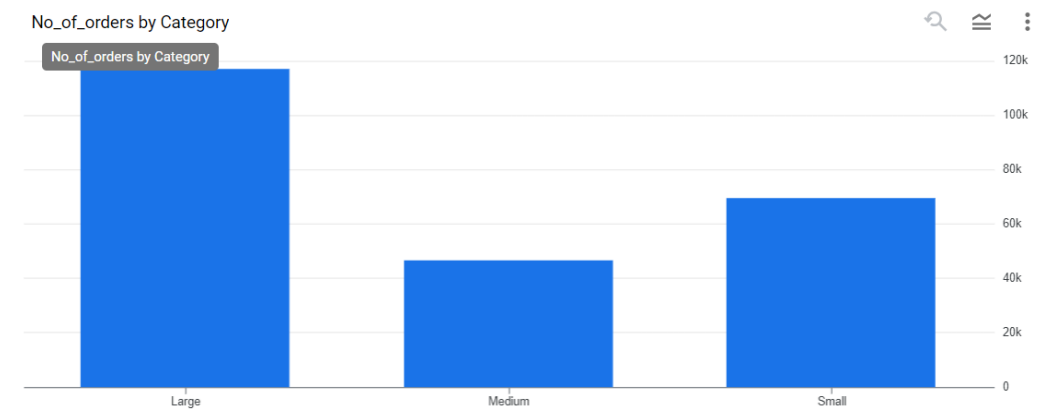
No_of_orders by Category



Question 2: Find the number of orders that are small, medium or large order value(small:0-5 dollars, medium:5-10 dollars, large:10+)

```
WITH cte AS
(
  SELECT BASKET_ID, SUM(SALES_VALUE) AS order_value
  FROM `ecommerce.transaction_data`
  GROUP BY BASKET_ID)
SELECT 'Small' AS Category, COUNT(*) AS No_of_orders
FROM cte
WHERE order_value < 5
UNION ALL
SELECT 'Medium' AS Category, COUNT(*) AS No_of_orders
FROM cte
WHERE order_value >= 5 AND order_value <= 10
UNION ALL
SELECT 'Large' AS Category, COUNT(*) AS No_of_orders
FROM cte
WHERE order_value > 10;
```

Row	Category	No_of_orders
1	Large	116941
2	Medium	46663
3	Small	69752



Question 3: Find top 3 stores with highest foot traffic for each week (Foot traffic: number of customers transacting)

```
with WeeklyFootTraffic as
(SELECT WEEK_NO, STORE_ID, COUNT(DISTINCT household_key) AS foot_traffic
FROM `ecommerce.transaction_data`
GROUP BY WEEK_NO, STORE_ID),
RankedStores AS
(SELECT WEEK_NO, STORE_ID, foot_traffic,
dense_rank() OVER (PARTITION BY WEEK_NO ORDER BY foot_traffic DESC) AS
store_rank
FROM WeeklyFootTraffic)
SELECT WEEK_NO, STORE_ID, foot_traffic
FROM RankedStores
WHERE store_rank <= 3
ORDER BY WEEK_NO, RankedStores.foot_traffic desc
```

Row	WEEK_NO	STORE_ID	foot_traffic
5	1	446	3
6	1	358	2
7	1	634	2
8	1	288	2
9	1	306	2
10	1	359	2
11	1	400	2
12	1	401	2
13	1	402	2
14	1	31642	2
15	1	412	2
16	1	427	2
17	1	443	2
18	2	32004	7
19	2	313	6
20	2	367	5
21	3	367	10
22	3	32004	9
23	3	356	8
24	4	367	17
25	4	32004	11
26	4	446	8

Question 4: Create a basic customer profiling with first, last visit, number of visits, average money spent per visit and total money spent order by highest avg money

```
select household_key,min(DAY) as first_visit, max(DAY) as last_visit,
count(distinct BASKET_ID) as no_of_visits, cast(round(sum(SALES_VALUE)) as
integer) as total_spends,
cast(round(sum(SALES_VALUE)/count(distinct BASKET_ID)) as integer) as
avg_spend
from `ecommerce.transaction_data`
group by household_key
order by 6 desc
```

Row	household_key	first_visit	last_visit	no_of_visits	total_spends	avg_spend
1	2042	52	683	26	2339	90
2	973	95	710	80	6876	86
3	1899	20	705	69	5790	84
4	1900	111	707	55	4228	77
5	1574	107	651	27	1843	68
6	2479	111	706	111	6955	63
7	1315	60	624	5	317	63
8	931	94	668	40	2455	61
9	1344	87	691	26	1570	60
10	248	29	704	53	3091	58
11	688	70	692	27	1559	58
12	1864	103	710	148	8537	58

Question 5: Do a single customer analysis selecting most spending customer for whom we have demographic information (because not all customers in transaction data are present in demographic table) (show the demographic as well as total spent)

```
WITH DemographicAndSpending AS
```

```
(SELECT d.household_key, d.AGE_DESC, d.MARITAL_STATUS_CODE, d.INCOME_DESC,
d.HOMEOWNER_DESC, d.HH_COMP_DESC, d.HOUSEHOLD_SIZE_DESC, d.KID_CATEGORY_DESC,
cast(round(SUM(t.SALES_VALUE)) as integer) AS total_spent
FROM `ecommerce.hh_demographic` d JOIN `ecommerce.transaction_data` t
ON d.household_key = t.household_key
GROUP BY d.household_key, d.AGE_DESC, d.MARITAL_STATUS_CODE,
d.INCOME_DESC, d.HOMEOWNER_DESC, d.HH_COMP_DESC, d.HOUSEHOLD_SIZE_DESC,
d.KID_CATEGORY_DESC)
```

```
SELECT household_key, AGE_DESC, MARITAL_STATUS_CODE, INCOME_DESC,
HOMEOWNER_DESC, HH_COMP_DESC, HOUSEHOLD_SIZE_DESC, KID_CATEGORY_DESC,
total_spent
FROM DemographicAndSpending
ORDER BY total_spent DESC
LIMIT 1;
```

Row	household_key	AGE_DESC	MARITAL_STATUS_CODE	INCOME_DESC	HOMEOWNER_DESC
1	1609	45-54	A	125-149K	Homeowner

HH_COMP_DESC	HOUSEHOLD_SIZE_DESC	KID_CATEGORY_DESC	total_spent
2 Adults Kids	5+	3+	13804

Question 6: Find products (product table : SUB_COMMODITY_DESC) which are most frequently bought together and the count of each combination bought together. do not print a combination twice (A-B / B-A)

```
select p1.SUB_COMMODITY_DESC as Product1, p2.SUB_COMMODITY_DESC as Product2,
count(*) as Pair_Count

from `ecommerce.transaction_data` t1 join `ecommerce.transaction_data` t2
on t1.BASKET_ID=t2.BASKET_ID
join `ecommerce.product` p1 on t1.PRODUCT_ID=p1.PRODUCT_ID
join `ecommerce.product` p2 on t2.PRODUCT_ID=p2.PRODUCT_ID
where p1.SUB_COMMODITY_DESC<p2.SUB_COMMODITY_DESC
group by p1.SUB_COMMODITY_DESC,p2.SUB_COMMODITY_DESC
order by 3 desc
```

Row	Product1	Product2	Pair_Count
1	FLUID MILK WHITE ONLY	YOGURT NOT MULTI-PACKS	5953
2	BANANAS	FLUID MILK WHITE ONLY	4365
3	FLUID MILK WHITE ONLY	SOFT DRINKS 12/18&15PK CA...	4326
4	FLUID MILK WHITE ONLY	MAINSTREAM WHITE BREAD	3934
5	BANANAS	YOGURT NOT MULTI-PACKS	3847
6	FLUID MILK WHITE ONLY	SHREDDED CHEESE	3840
7	FLUID MILK WHITE ONLY	SFT DRNK 2 LITER BTL CARB I...	3494
8	FRZN SS PREMIUM ENTREES/...	YOGURT NOT MULTI-PACKS	3344
9	BABY FOOD - BEGINNER	BABY FOOD JUNIOR ALL BRAN...	3290
10	SHREDDED CHEESE	YOGURT NOT MULTI-PACKS	3189

Question 7: Find the weekly change in Revenue Per Account (RPA) (difference in spending by each customer compared to last week)(use lag function)

```
with cte as
(select household_key, WEEK_NO, sum(SALES_VALUE) as week_rev
from `ecommerce.transaction_data`
group by household_key, WEEK_NO)

select household_key, WEEK_NO, round(tbl.week_rev,2) as week_rev,
round(tbl.prev_rev,2) as prev_week_rev, round(tbl.week_rev-tbl.prev_rev,2)
as weekly_rev_change
from
(select household_key, WEEK_NO, cte.week_rev,
lag(cte.week_rev) over(partition by household_key order by WEEK_NO) as
prev_rev
from cte) tbl
order by 1,2
```

Row	household_key	WEEK_NO	week_rev	prev_week_rev	weekly_rev_change
1	1	8	42.58	null	null
2	1	10	14.01	42.58	-28.57
3	1	13	14.03	14.01	0.02
4	1	14	25.71	14.03	11.68
5	1	15	10.98	25.71	-14.73
6	1	16	9.09	10.98	-1.89
7	1	17	13.98	9.09	4.89
8	1	19	47.35	13.98	33.37
9	1	20	31.77	47.35	-15.58
10	1	22	38.98	31.77	7.21
11	1	23	26.36	38.98	-12.62
12	1	24	35.13	26.36	8.77
13	1	25	17.23	35.13	-17.9
14	1	26	32.06	17.23	14.83
15	1	28	42.34	32.06	10.28
16	1	30	18.43	42.34	-23.91

Additional Queries

Q1: Customer Retention Rate (Customers Returning Within 4 Weeks)

Customer Retention Rate, especially tracking customers returning within 4 weeks, is crucial for understanding loyalty and engagement. High retention indicates satisfaction and effective marketing, while low retention signals potential issues that need addressing to maintain steady revenue and long-term profitability.

```
WITH first_visits AS
(
  SELECT household_key, MIN(WEEK_NO) AS first_week
  FROM `ecommerce.transaction_data`
  GROUP BY household_key),
subsequent_visits AS
(
  SELECT household_key, MIN(WEEK_NO) AS next_visit_week
  FROM `ecommerce.transaction_data` as t
  WHERE WEEK_NO > (SELECT first_week FROM first_visits WHERE
    first_visits.household_key = t.household_key) + 4
  GROUP BY household_key)
SELECT COUNT(*) AS total_customers,
SUM(CASE WHEN next_visit_week IS NOT NULL THEN 1 ELSE 0 END) AS
retained_customers,
SUM(CASE WHEN next_visit_week IS NOT NULL THEN 1 ELSE 0 END) / COUNT(*) *
100 AS retention_rate
FROM first_visits
LEFT JOIN subsequent_visits ON first_visits.household_key =
subsequent_visits.household_key;
```

Row	total_customers	retained_customers	retention_rate
1	2500	2491	99.64

Q2. Total Sales by Product Category

Total sales by product category is essential for understanding revenue drivers, customer preferences, and product performance. It helps businesses optimize inventory, tailor marketing strategies, and make informed decisions on resource allocation and product development.

```
select p.DEPARTMENT, cast(round(sum(t.SALES_VALUE)) as integer) as
Total_Sales
from `ecommerce.transaction_data` t join `ecommerce.product` p
on t.PRODUCT_ID=p.PRODUCT_ID
group by p.DEPARTMENT
order by 2 desc
```

Row	DEPARTMENT	Total_Sales
1	GROCERY	2046695
2	DRUG GM	527589
3	PRODUCE	279720
4	MEAT	274036
5	KIOSK-GAS	269462
6	MEAT-PCKGD	206492
7	DELI	130322
8	MISC SALES TRAN	62634
9	PASTRY	61787
10	NUTRITION	48840

Q3. Average Basket Size by Store

Average basket size by store is crucial for analysing sales performance, customer buying behaviour, and store efficiency. It helps identify opportunities for increasing revenue per transaction and optimizing product placement and promotions within each store.

```
Select STORE_ID, avg(QUANTITY) as avg_basket_size
from `ecommerce.transaction_data`
group by STORE_ID
order by 2 desc
```

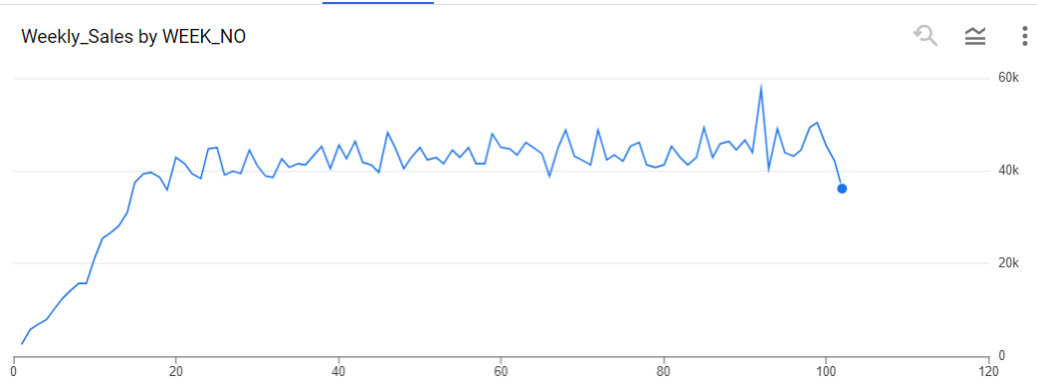
Row	STORE_ID	avg_basket_size
1	3065	24007.0
2	489	22842.0
3	3098	17956.0
4	144	14833.0
5	896	14499.0
6	648	10802.0
7	2790	10762.0
8	84	10007.0
9	3006	9857.0
10	3090	9436.5

Q4. Weekly Sales Trend

Weekly sales trend analysis is crucial for understanding fluctuations in consumer demand, identifying seasonality, and evaluating the effectiveness of promotions or marketing campaigns. It helps businesses adjust strategies in real-time to maximize revenue and optimize inventory management.

```
Select WEEK_NO, round(sum(SALES_VALUE),2) as Weekly_Sales
from `ecommerce.transaction_data`
group by WEEK_NO
order by 1
```

Row	WEEK_NO	Weekly_Sales
1	1	2453.52
2	2	5854.47
3	3	6822.98
4	4	7915.42
5	5	10038.67
6	6	12566.74
7	7	14041.61
8	8	15679.87
9	9	15653.89
10	10	21458.3



Q5. Identify the top 5 pairs of product categories that are most frequently purchased together by the same household within the same basket.

Understanding which product categories are frequently purchased together can reveal important cross-selling opportunities. For example, if certain categories like “Groceries” and “Drug” are often bought together, the business could create bundle offers, adjust store layouts, or tailor marketing campaigns to encourage these purchasing behaviours further, thus driving higher sales and improving the shopping experience.

```
WITH department_pairs AS
(
  SELECT t1.household_key, t1.BASKET_ID,
  LEAST(p1.DEPARTMENT, p2.DEPARTMENT) AS Department_A,
  GREATEST(p1.DEPARTMENT, p2.DEPARTMENT) AS Department_B
  FROM `ecommerce.transaction_data` t1 JOIN
  `ecommerce.transaction_data` t2 ON t1.household_key = t2.household_key
  AND t1.BASKET_ID = t2.BASKET_ID AND t1.PRODUCT_ID < t2.PRODUCT_ID
  JOIN `ecommerce.product` p1 ON t1.PRODUCT_ID = p1.PRODUCT_ID
  JOIN `ecommerce.product` p2 ON t2.PRODUCT_ID = p2.PRODUCT_ID
  WHERE p1.DEPARTMENT <> p2.DEPARTMENT)
SELECT Department_A, Department_B, COUNT(*) AS Pair_Count
FROM department_pairs
GROUP BY Department_A, Department_B
ORDER BY Pair_Count DESC
LIMIT 5;
```

Row	Department_A	Department_B	Pair_Count
1	GROCERY	PRODUCE	1014624
2	DRUG GM	GROCERY	763097
3	GROCERY	MEAT-PCKGD	515318
4	GROCERY	MEAT	373868
5	DELI	GROCERY	214212

INSIGHTS:

1. Total Sales- 4029338
2. Total Number of Orders-233356
3. Average Order Value- 17
4. Total number of products- 2383
5. Total number of households- 801
6. Number of Orders segmentation w.r.t order value-
 - a. Small (0-10 dollars): 115045
 - b. Medium (10-20 dollars): 51000
 - c. Large (20+ dollars): 67311
7. Number of Orders segmentation w.r.t order value-
 - a. Small (0-5 dollars): 69752
 - b. Medium (5-10 dollars): 46663
 - c. Large (10+ dollars): 116941
8. Identified for each week top 3 stores with highest foot traffic
9. Created a list of basic customer profiling with first, last visit, number of visits, average money spent per visit and total money spent. With household key 2042 as example with highest average spent of 90 dollars
 - a. First visit: week no. 52
 - b. Last visit: 683
 - c. Total spend: 2339
10. Household_key 1609 is the most spending customer with total spends of 13804 dollars and following demographic details:
 - a. Age Group: 45-54
 - b. Marital Status: Married
 - c. Income group: 125k-149k
 - d. Home Ownership status: Home owner
 - e. Household size: 5+
 - f. Family composition: 2 Adult kids
 - g. Number of kids: 3+
11. Products pair that are most frequently bought together are:
 - a. Fluid Milk White Only and Yogurt Not Multi-Packs
 - b. Bananas and Fluid Milk White Only
12. Found the weekly change in Revenue Per Account (RPA) (difference in spending by each customer compared to last week)
13. Identified Customer Retention Rate (Customers Returning Within 4 Weeks) as 99.64%
14. Grocery is the most selling product category in terms of sales value followed by Drug GM and Produce
15. Store_ID 3065 has the highest average Basket size followed by 489 and 3098
16. Studied weekly sales trend as shared in detail above
17. Identified the top pairs of product categories that are most frequently purchased together by the same household within the same basket.
 - a. Grocery and Produce
 - b. Grocery and Drug GM

Recommendations

1. **Enhance Cross-Selling Strategies:**
 - Leverage the identified product pairs frequently bought together (e.g., Fluid Milk and Yogurt) by creating bundled offers or promotions. This could be particularly effective in boosting average basket size and overall sales.
 - Consider strategic placement of these products in both physical and online stores to encourage impulse purchases and increase customer convenience.
2. **Focus on High-Value Customer Segments:**
 - Based on the customer profiling data, high-spending households such as household_key 1609 should be targeted with personalized offers, loyalty programs, and premium services. Tailoring marketing campaigns to high-income, large households with children can lead to higher customer satisfaction and increased lifetime value.
 - Enhance customer retention strategies for these segments by offering incentives for frequent shopping and exclusive benefits for repeat customers.
3. **Optimize Product Offerings in Key Categories:**
 - With Grocery, Drug GM, and Produce being the top-selling categories, ensure these departments are well-stocked, priced competitively, and frequently refreshed with new or seasonal products.
 - Consider expanding product lines in these categories based on emerging trends and customer preferences to capture a larger share of wallet.
4. **Improve Store-Specific Strategies:**
 - For stores with the highest average basket size (e.g., Store_ID 3065), analyze the factors contributing to this success and replicate these strategies across other stores. This could include store layout, product placement, or localized marketing efforts.
 - Implement training and performance incentives for store staff in locations with high foot traffic to ensure customer satisfaction and maximize sales during peak periods.
5. **Leverage Customer Retention Insights:**
 - With a high customer retention rate (99.64% returning within 4 weeks), continue to nurture customer loyalty through regular engagement, personalized communication, and rewards for frequent purchases.
 - Introduce a referral program to leverage loyal customers in acquiring new customers, further boosting retention and sales.
6. **Monitor and Adapt to Weekly Sales Trends:**
 - Use weekly sales trend data to optimize inventory management, ensuring high-demand products are always available while minimizing overstock. This will help in reducing waste and maximizing profitability.
 - Adapt marketing and promotional strategies based on identified seasonal trends or sales spikes to capture maximum revenue during peak periods.
7. **Expand Customer Insights with Advanced Analytics:**
 - Consider implementing more advanced analytics, such as predictive modelling, to anticipate future sales trends, customer behavior, and product performance. This proactive approach can help in staying ahead of market changes and customer expectations.

By implementing these recommendations, the e-commerce company can drive more targeted growth, improve customer satisfaction, and maximize profitability across all areas of the business.