E-commerce SQL analysis

Problem Statement

Analysing the sales, product, and customer data for an e-commerce company. getting various insights and calculating various KPI and data with SQL in Big Query.

Dataset:

https://drive.google.com/drive/folders/1xU91jKUknFRtBlC9vrKUSfhSC9_kvb30

Data Disctionary:

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+

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

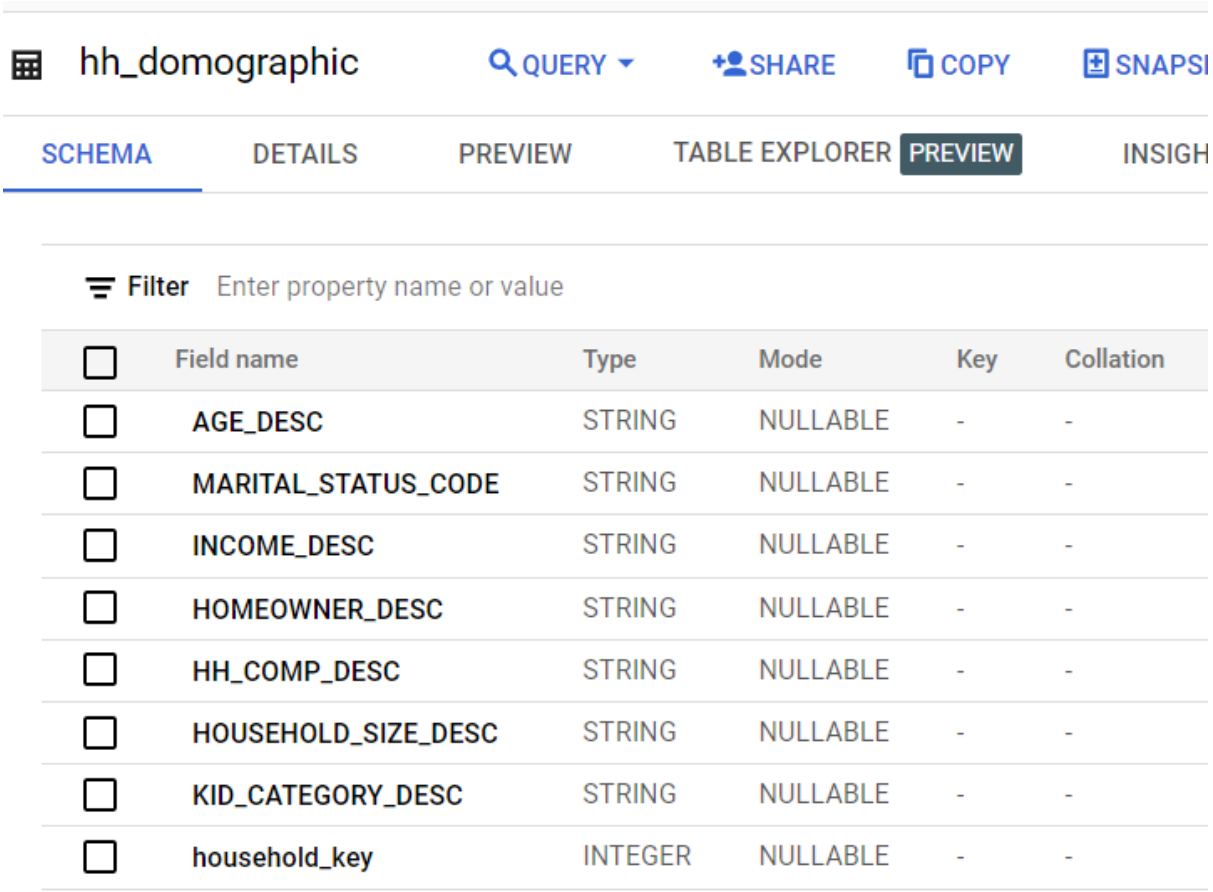
Variable	Description
PRODUCT_ID	Number that uniquely identifies each product
DEPARTMENT	Groups similar products together
COMMODITY_DESC	Groups similar products together at a lower level
SUB_COMMODITY_DESC	Groups similar products together at the lowest level
MANUFACTURER	Code that links products with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

Goal:

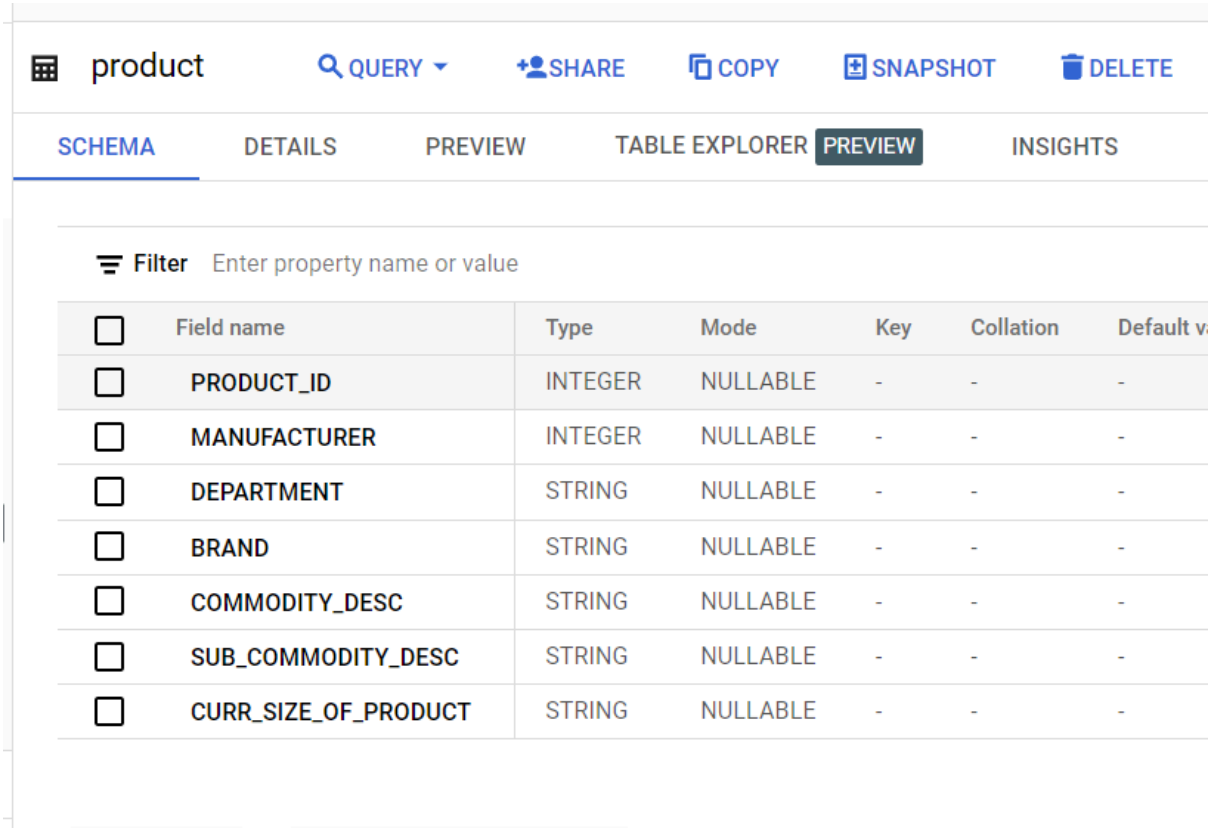
This project aims to leverage the power of e-commerce data (sales, product, and demographic(Customer)) analysed through SQL to unlock actionable insights driving profitable growth. By delving into customer behaviour, product trends, and sales

patterns, we will uncover hidden value that can inform key business decisions. We need to find these patterns and calculate various metrics and KPIs that suit the data and the goal.

Table description:-



Field name	Type	Mode	Key	Collation
AGE_DESC	STRING	NULLABLE	-	-
MARITAL_STATUS_CODE	STRING	NULLABLE	-	-
INCOME_DESC	STRING	NULLABLE	-	-
HOMEOWNER_DESC	STRING	NULLABLE	-	-
HH_COMP_DESC	STRING	NULLABLE	-	-
HOUSEHOLD_SIZE_DESC	STRING	NULLABLE	-	-
KID_CATEGORY_DESC	STRING	NULLABLE	-	-
household_key	INTEGER	NULLABLE	-	-



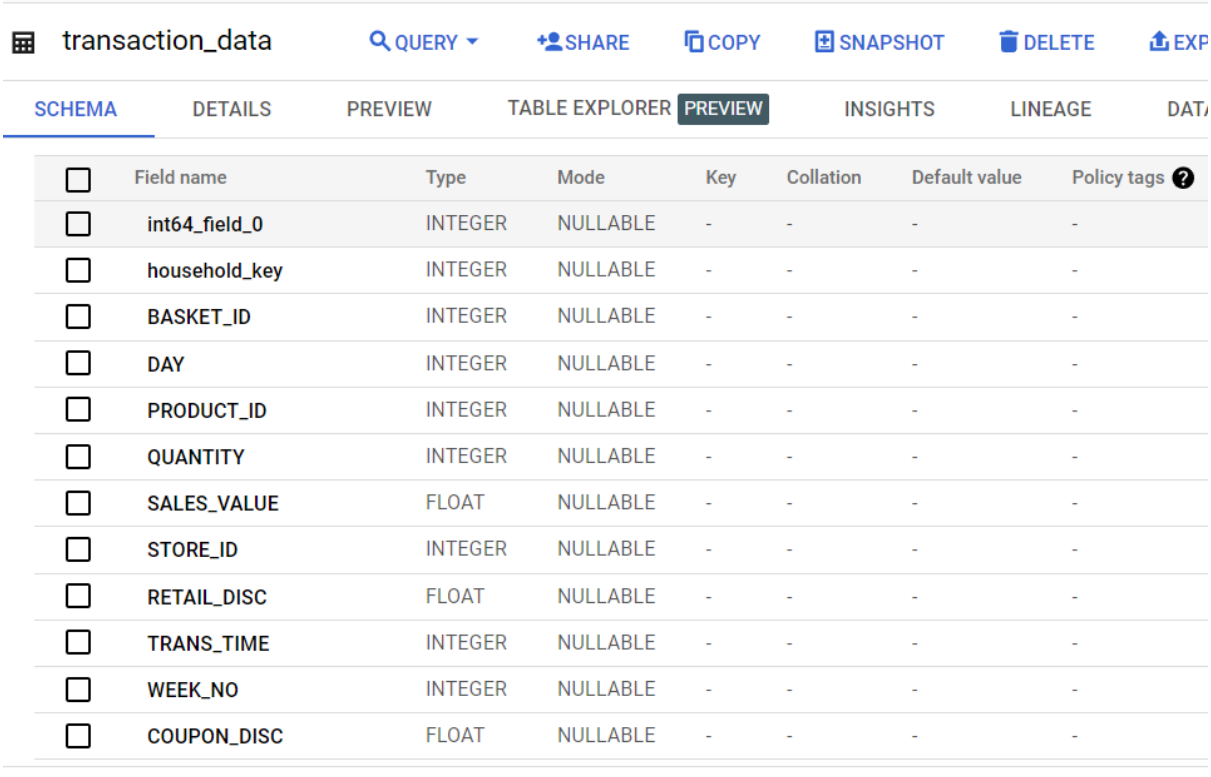
product

QUERY SHARE COPY SNAPSHOT DELETE

SCHEMA DETAILS PREVIEW TABLE EXPLORER PREVIEW INSIGHTS

Filter Enter property name or value

Field name	Type	Mode	Key	Collation	Default v
PRODUCT_ID	INTEGER	NULLABLE	-	-	-
MANUFACTURER	INTEGER	NULLABLE	-	-	-
DEPARTMENT	STRING	NULLABLE	-	-	-
BRAND	STRING	NULLABLE	-	-	-
COMMODITY_DESC	STRING	NULLABLE	-	-	-
SUB_COMMODITY_DESC	STRING	NULLABLE	-	-	-
CURR_SIZE_OF_PRODUCT	STRING	NULLABLE	-	-	-



transaction_data

QUERY SHARE COPY SNAPSHOT DELETE EXP

SCHEMA DETAILS PREVIEW TABLE EXPLORER PREVIEW INSIGHTS LINEAGE DAT

Field name	Type	Mode	Key	Collation	Default value	Policy tags
int64_field_0	INTEGER	NULLABLE	-	-	-	-
household_key	INTEGER	NULLABLE	-	-	-	-
BASKET_ID	INTEGER	NULLABLE	-	-	-	-
DAY	INTEGER	NULLABLE	-	-	-	-
PRODUCT_ID	INTEGER	NULLABLE	-	-	-	-
QUANTITY	INTEGER	NULLABLE	-	-	-	-
SALES_VALUE	FLOAT	NULLABLE	-	-	-	-
STORE_ID	INTEGER	NULLABLE	-	-	-	-
RETAIL_DISC	FLOAT	NULLABLE	-	-	-	-
TRANS_TIME	INTEGER	NULLABLE	-	-	-	-
WEEK_NO	INTEGER	NULLABLE	-	-	-	-
COUPON_DISC	FLOAT	NULLABLE	-	-	-	-

Here are some of the insights that have been derived from writing SQL query from the data.

1. **The number of orders that have small, medium or large order value (small:0-10 dollars, medium:10-20 dollars, large:20+)**

```sql
SELECT
  CASE
    WHEN SALES_VALUE BETWEEN 0 AND 10 THEN 'Small'
    WHEN SALES_VALUE BETWEEN 10 AND 20 THEN 'Medium'
    ELSE 'Large'
  END AS Order_Size,
  COUNT(*) AS Number_of_Orders
FROM `data_analytics.transaction_data`
GROUP BY Order_Size
```

Query results

JOB INFORMATION | RESULTS | CHART | JSON

Row	Order_Size	Number_of_Orders
1	Small	1016841
2	Medium	21653
3	Large	10081

2. **Top 3 stores with highest foot traffic for each week**

```sql
WITH Store_Traffic AS (
  SELECT STORE_ID, WEEK_NO, COUNT(DISTINCT HOUSEHOLD_KEY) AS Customer_Count
  FROM `data_analytics.transaction_data`
  GROUP BY STORE_ID, WEEK_NO
)
SELECT STORE_ID, WEEK_NO, Customer_Count
FROM (
  SELECT *, ROW_NUMBER() OVER (PARTITION BY WEEK_NO ORDER BY Customer_Count
DESC) AS Rank
  FROM Store_Traffic
) AS Ranked_Stores
WHERE Rank <= 3
```

Query results

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS

Row	STORE_ID	WEEK_NO	Customer_Count
1	32004	6	16
2	367	6	13
3	335	6	10
4	32004	59	29

3. Basic customer profiling.

```
SELECT HOUSEHOLD_KEY,
       MIN(DAY) AS First_Visit,
       MAX(DAY) AS Last_Visit,
       COUNT(BASKET_ID) AS Number_of_Visits,
       SUM(SALES_VALUE) AS Total_Spent,
       AVG(SALES_VALUE) AS Avg_Spent_Per_Visit
FROM `data_analytics.transaction_data`
GROUP BY HOUSEHOLD_KEY
ORDER BY Avg_Spent_Per_Visit DESC;
```

Query results

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	HOUSEHOLD_KEY	First_Visit	Last_Visit	Number_of_Visits	Total_Spent	Avg_Spent_Per_Visit
1	1730	34	707	99	1656.760000000...	16.73494949494...
2	1727	109	118	9	114.51	12.72333333333...
3	1339	52	701	18	187.53	10.41833333333...
4	991	44	665	44	451.6	10.26363636363...

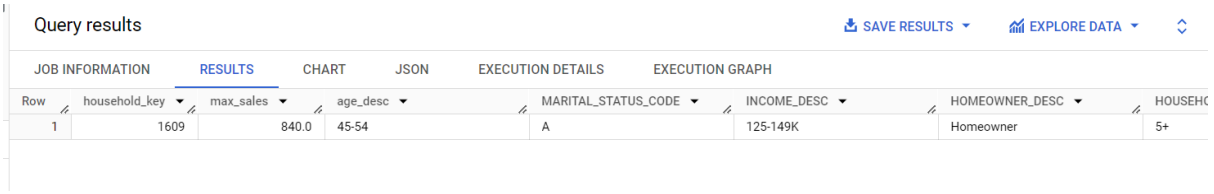
Results per page: 50

4. Single customer analysis (highest spender)

```
with maxsales_per_household as (select household_key, max(SALES_VALUE) as
max_sales
from `data_analytics.transaction_data`
group by household_key)


select
m.household_key,m.max_sales,h.age_desc,h.MARITAL_STATUS_CODE,h.INCOME_DESC,h
.HOMEOWNER_DESC,h.HOUSEHOLD_SIZE_DESC,h.INCOME_DESC,h.KID_CATEGORY_DESC
```

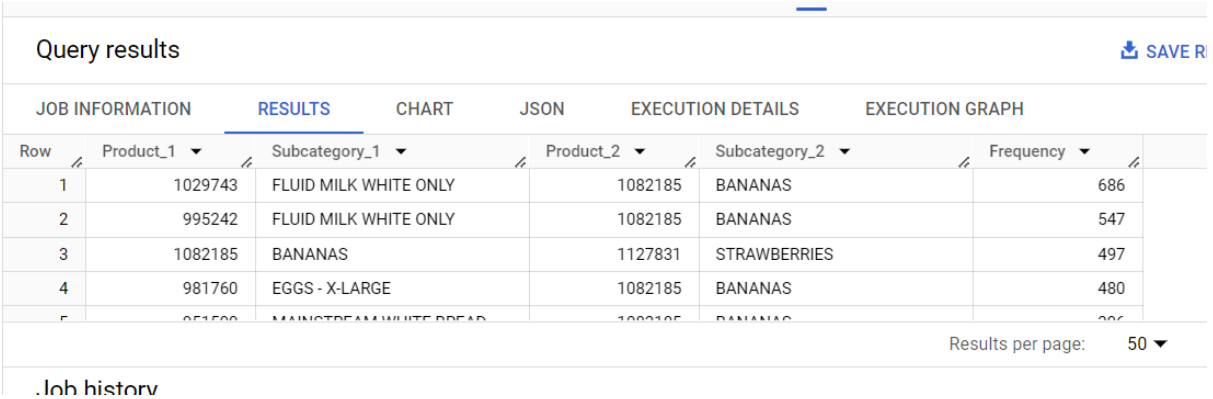
```
from maxsales_per_household m
join `data_analytics.hh_domographic` h
on m.household_key=h.household_key
order by m.max_sales DESC
limit 1
```



Row	household_key	max_sales	age_desc	MARITAL_STATUS_CODE	INCOME_DESC	HOMEOWNER_DESC	HOUSEHO
1	1609	840.0	45-54	A	125-149K	Homeowner	5+

5. Frequently bought together products

```
WITH Product_Pairs AS (
    SELECT a.BASKET_ID,
           LEAST(a.PRODUCT_ID, b.PRODUCT_ID) AS Product_1,
           GREATEST(a.PRODUCT_ID, b.PRODUCT_ID) AS Product_2
    FROM `data_analytics.transaction_data` a
    JOIN `data_analytics.transaction_data` b ON a.BASKET_ID = b.BASKET_ID
    WHERE a.PRODUCT_ID < b.PRODUCT_ID
),
Product_Frequency AS (
    SELECT Product_1, Product_2, COUNT(*) AS Frequency
    FROM Product_Pairs
    GROUP BY Product_1, Product_2
)
SELECT pf.Product_1, p1.SUB_COMMODITY_DESC AS Subcategory_1,
       pf.Product_2, p2.SUB_COMMODITY_DESC AS Subcategory_2,
       pf.Frequency
FROM Product_Frequency pf
JOIN `data_analytics.product` p1 ON pf.Product_1 = p1.PRODUCT_ID
JOIN `data_analytics.product` p2 ON pf.Product_2 = p2.PRODUCT_ID
ORDER BY pf.Frequency DESC
```

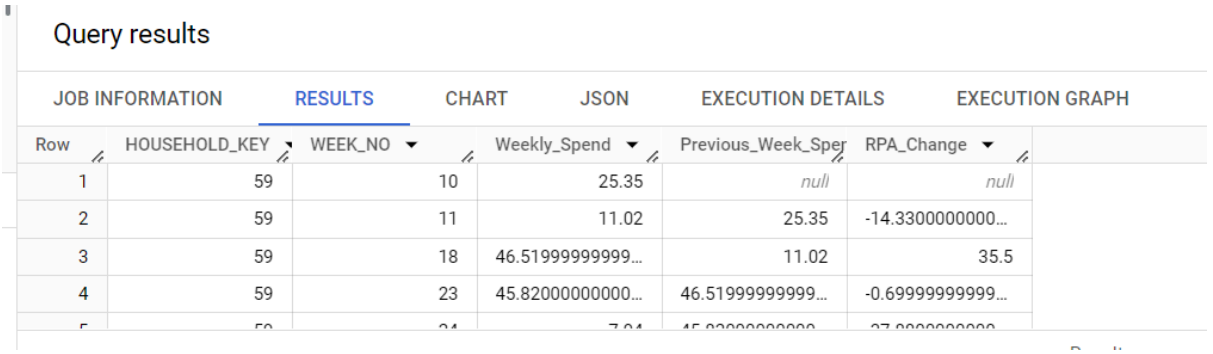


Row	Product_1	Subcategory_1	Product_2	Subcategory_2	Frequency
1	1029743	FLUID MILK WHITE ONLY	1082185	BANANAS	686
2	995242	FLUID MILK WHITE ONLY	1082185	BANANAS	547
3	1082185	BANANAS	1127831	STRAWBERRIES	497
4	981760	EGGS - X-LARGE	1082185	BANANAS	480

6. Weekly change in Revenue Per Account (RPA)

```
WITH Weekly_Revenue AS (
```

```
  SELECT HOUSEHOLD_KEY, WEEK_NO, SUM(SALES_VALUE) AS Weekly_Spend
  FROM `data_analytics.transaction_data`
  GROUP BY HOUSEHOLD_KEY, WEEK_NO
)
SELECT HOUSEHOLD_KEY, WEEK_NO,
       Weekly_Spend,
       LAG(Weekly_Spend, 1) OVER (PARTITION BY HOUSEHOLD_KEY ORDER BY
WEEK_NO) AS Previous_Week_Spend,
       (Weekly_Spend - LAG(Weekly_Spend, 1) OVER (PARTITION BY HOUSEHOLD_KEY
ORDER BY WEEK_NO)) AS RPA_Change
FROM Weekly_Revenue;
```



Query results

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	HOUSEHOLD_KEY	WEEK_NO	Weekly_Spend	Previous_Week_Spe	RPA_Change
1	59	10	25.35	*null*	*null*
2	59	11	11.02	25.35	-14.3300000000...
3	59	18	46.51999999999...	11.02	35.5
4	59	23	45.82000000000...	46.51999999999...	-0.69999999999...

7. **Top 10 products based on the total sales**

```
SELECT
  p.SUB_COMMODITY_DESC,
  SUM(t.sales_value) AS total_sales
FROM `data_analytics.transaction_data` t
JOIN `data_analytics.product` p ON t.product_id = p.product_id
GROUP BY p.SUB_COMMODITY_DESC
ORDER BY total_sales DESC
LIMIT 10;
```

Query results

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DE

Row	SUB_COMMODITY_DESC	total_sales
1	GASOLINE-REG UNLEADED	255862.0500000...
2	SOFT DRINKS 12/18&15PK CA...	65828.03999999...
3	FLUID MILK WHITE ONLY	64291.17999999...
4	BEERALEMALT LIQUORS	62697.95999999...
5	CIGARETTES	39761.87000000...
6	CHOICE BEEF	30455.36000000...
7	SHREDDED CHEESE	27515.16999999...
8	PRIMAL	25876.27000000...
9	PREMIUM	25518.25000000...
10	TOILET TISSUE	24574.22000000...

8. **demographic groups (age, income, household size) have the highest average order value**

```
SELECT
    d.AGE_DESC,
    d.INCOME_DESC,
    d.HOUSEHOLD_SIZE_DESC,
    AVG(t.SALES_VALUE) AS Average_Order_Value
FROM `data_analytics.transaction_data` t
JOIN `data_analytics.hh_domographic` d ON t.HOUSEHOLD_KEY = d.HOUSEHOLD_KEY
GROUP BY d.AGE_DESC, d.INCOME_DESC, d.HOUSEHOLD_SIZE_DESC
ORDER BY Average_Order_Value DESC;
```

Query results

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	AGE_DESC	INCOME_DESC	HOUSEHOLD_SIZE_DESC	Average_Order_Value
1	45-54	Under 15K	4	6.230102915951...
2	25-34	175-199K	1	5.594245129870...
3	65+	Under 15K	2	5.045910852713...
4	35-44	175-199K	2	4.985432432432...

Results per page:

9. **Which product categories have the highest total sales?**

```
SELECT p.COMMODITY_DESC, SUM(t.SALES_VALUE) AS Total_Sales
FROM `data_analytics.transaction_data` t
JOIN `data_analytics.product` p ON t.PRODUCT_ID = p.PRODUCT_ID
GROUP BY p.COMMODITY_DESC
ORDER BY Total_Sales DESC;
```

Query results

JOB INFORMATION | RESULTS | CHART | JSON

Row	COMMODITY_DESC	Total_Sales
1	COUPON/MISC ITEMS	258796.8800000...
2	SOFT DRINKS	136621.9999999...
3	BEEF	125580.6900000...
4	FLUID MILK PRODUCTS	81817.18000000...
5	CHEESE	75994.26000000...

10. **What is the repeat purchase rate for each demographic group?**

```
SELECT d.AGE_DESC, d.INCOME_DESC, d.HOUSEHOLD_SIZE_DESC,
       COUNT(DISTINCT t.HOUSEHOLD_KEY) / COUNT(t.HOUSEHOLD_KEY) AS
Repeat_Purchase_Rate
FROM `data_analytics.transaction_data` t
JOIN `data_analytics.hh_domographic` d ON t.HOUSEHOLD_KEY = d.HOUSEHOLD_KEY
GROUP BY d.AGE_DESC, d.INCOME_DESC, d.HOUSEHOLD_SIZE_DESC
ORDER BY Repeat_Purchase_Rate DESC
```

Query results

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTION GRAPH

Row	AGE_DESC	INCOME_DESC	HOUSEHOLD_SIZE_DESC	Repeat_Purchase_Ra
1	19-24	15-24K	3	0.005681818181...
2	45-54	150-174K	3	0.005291005291...
3	35-44	15-24K	5+	0.005181347150...
4	45-54	250K+	2	0.004291845493...
5	45-54	15-24K	5+	0.004065040650...

Results per page:

11. **Which stores have the highest average basket size?**

```
SELECT t.STORE_ID, AVG(t.SALES_VALUE) AS Average_Basket_Size
FROM `data_analytics.transaction_data` t
GROUP BY t.STORE_ID
```

```
ORDER BY Average_Basket_Size DESC;
```

JOB INFORMATION | RESULTS | CHART

Row	STORE_ID	Average_Basket_Size
1	3098	43.62
2	3422	39.99
3	144	39.0
4	896	30.0
5	640	30.42

12. Which products generate the most discount usage?

```
SELECT p.PRODUCT_ID, p.SUB_COMMODITY_DESC,
       SUM(t.COUPON_DISC + t.RETAIL_DISC) AS Total_Discounts
FROM `data_analytics.transaction_data` t
JOIN `data_analytics.product` p ON t.PRODUCT_ID = p.PRODUCT_ID
GROUP BY p.PRODUCT_ID, p.SUB_COMMODITY_DESC
ORDER BY Total_Discounts DESC
```

Query results

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS

Row	PRODUCT_ID	SUB_COMMODITY_DESC	Total_Discounts
1	1033887	CAN CATFD GOURMET/SUP P...	0.01
2	995436	CORNISH HEN	0.0
3	861973	SFT DRNK MLT-PK BTL CARB (...	0.0
4	1124520	CIGARETTES	0.0
5	995997	MEAT-HAM BULK	0.0

Insights:-

- **Order Size Distribution:** The analysis revealed different order size categories—small, medium, and large—based on sales value. This classification helps understand the customer spending habits and the relative distribution of order values.
- **Store Foot Traffic:** The top 3 stores with the highest foot traffic for each week were identified. This information can be used to focus marketing efforts and inventory management on high-traffic stores, ensuring they are well-stocked and efficiently managed.
- **Customer Profiling:** Basic customer profiling, including first and last visits, total spending, and the average spent per visit, helps tailor personalized marketing and customer retention strategies.
- **High-Value Customers:** The highest spenders were identified through a single customer analysis. These customers can be targeted for loyalty programs and exclusive offers to further enhance their engagement and satisfaction.
- **Product Bundling:** Frequently bought together products were identified, offering insights into effective product bundling strategies. This can boost sales by promoting frequently paired items together as discounts or combo deals.
- **Weekly Revenue Fluctuations:** The analysis of revenue per account (RPA) highlighted week-over-week changes, which provides a deeper understanding of customer purchasing patterns and can help optimize promotions or offers during specific periods.
- **Top-Selling Products:** Identifying the top 10 products based on total sales highlights the most popular products, which can inform inventory management and marketing priorities.
- **Demographic Insights:** Insights into which demographic groups (age, income, household size) have the highest average order value offer a clear direction for demographic-specific marketing and product recommendations.
- **Product Category Performance:** Analysis of product categories with the highest total sales highlights which categories drive the most revenue, allowing for better category management and marketing focus.
- **Repeat Purchase Rates:** Understanding which demographic groups have the highest repeat purchase rates can help enhance retention strategies and target campaigns aimed at increasing customer loyalty.
- **Store Basket Size:** Stores with the highest average basket size were identified. This insight helps in understanding which locations encourage larger purchases and how to replicate those strategies across other stores.
- **Discount Usage:** Insights into which products generated the most discount usage can inform future promotions and discount strategies, ensuring that high-demand items are leveraged effectively during sales events.

Recommendations:

- **Personalized Marketing**: Utilise customer profiling and demographic insights to create personalised marketing campaigns. Target high-value and high-frequency customers with tailored offers to increase engagement and loyalty.
- **Inventory and Store Management**: Focus inventory replenishment on the top stores with the highest traffic and largest average basket sizes. This will ensure these stores are well-stocked and can continue driving high sales.
- **Product Bundling**: Implement product bundling strategies based on the frequently bought together analysis. Offering discounts or special promotions for these product pairs can drive higher transaction values.
- **Loyalty Programs for High-Spenders**: Introduce or enhance loyalty programs targeting high-value customers, offering exclusive discounts, early access to sales, or other perks to encourage continued engagement.
- **Optimise Promotional Strategies**: Use the analysis of weekly changes in revenue and discount usage to optimise promotional efforts. Target periods of low revenue with special deals or highlight high-discount products in promotional campaigns.
- **Customer Retention Initiatives**: Focus on increasing the repeat purchase rate among key demographic groups. Utilise insights from demographic data to tailor retention strategies like personalised offers or post-purchase follow-ups.
- **Category-Specific Campaigns**: Highlight the product categories that generate the most sales in marketing campaigns, ensuring that they remain well-promoted and continue driving revenue growth.