

E-commerce SQL Analysis

- 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,
count(distinct BASKET_ID) as Number_of_orders,
round(sum(SALES_VALUE),2) as Total_sale
from `ecommerce.transaction`
group by 1)

select
count(Number_of_orders) as Total_orders,
case when total_sale >= 0 and total_sale < 10 then "small"
when total_sale >= 10 and total_sale < 20 then "medium"
when total_sale >= 20 then "large"
end as Order_size
from cte
group by 2
```

Row	Total_orders	Order_size
1	68422	large
2	115045	small
3	49889	medium

Insights : Total Count for the order which are > 0 and less than 10 is 68422 which are considered as large Orders. Count for orders which are >= 10 and less than 20 is 115045 which is considered as small and count for > 20 is 49889 which is considered as medium.

- 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,
count(distinct BASKET_ID) as Number_of_orders,
round(sum(SALES_VALUE),2) as Total_sale
from `ecommerce.transaction`
group by 1)

select
count(Number_of_orders) as Total_orders,
case when total_sale >= 0 and total_sale < 5 then "small"
when total_sale >= 5 and total_sale < 10 then "medium"
```

```

when total_sale >= 10 then "large"
end as Order_size
from cte
group by 2

```

Row	Total_orders	Order_size
1	118311	large
2	69752	small
3	45293	medium

Insights : Total Count for the order which are > 0 and less than 5 is 118311 which are considered as large Orders. Count for orders which are >= 5 and less than 10 is 69752 which is considered as small and count for >= 10 is 45293 which is considered as medium.

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

```

with cte as
(select
count(BASKET_ID) as basket_cnt, WEEK_NO, STORE_ID,
dense_rank() over (partition by WEEK_NO order by count(BASKET_ID)
desc) as rnk
from `ecommerce.transaction`
group by 2,3)

```

```

select WEEK_NO, rnk as POSITION, STORE_ID from cte where rnk <=3
order by WEEK_NO, POSITION

```

Row	WEEK_NO	POSITION	STORE_ID
1	1	1	324
2	1	2	321
3	1	3	32004
4	2	1	375
5	2	2	292
6	2	3	315
7	3	1	367
8	3	2	375
9	3	3	356
10	4	1	367
11	4	2	375

- 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

```
WITH cte1 AS (
    SELECT
        household_key,
        SALES_VALUE,
        LPAD(CAST(DAY AS STRING), 3, '0') AS day,
        LPAD(CAST(TRANS_TIME AS STRING), 4, '0') AS trans_time
    FROM
        `ecommerce.transaction`
),
cte2 AS (
    SELECT
        household_key,
        COUNT(*) AS number_of_visits,
        MIN(CONCAT(day, "_", LEFT(trans_time, 2), ":",
RIGHT(trans_time, 2))) AS first_visit,
        MAX(CONCAT(day, "_", LEFT(trans_time, 2), ":",
RIGHT(trans_time, 2))) AS last_visit,
        ROUND(SUM(SALES_VALUE), 3) AS total_sales,
        ROUND(AVG(SALES_VALUE), 3) AS avg_sales
    FROM
        cte1
    GROUP BY
        household_key
)

SELECT * FROM cte2;
```

Row	household_key	number_of_visits	first_visit	last_visit	total_sales	avg_sales
1	7	643	023_14:35	709_18:49	1733.88	2.697
2	27	1245	035_18:29	711_20:59	2814.89	2.261
3	32	492	049_18:59	711_10:46	1674.1	3.403
4	54	17	034_23:15	495_15:05	50.26	2.956
5	91	410	056_01:24	710_11:03	1355.99	3.307
6	98	671	001_19:37	710_21:12	1250.37	1.863
7	111	299	013_16:41	687_13:28	714.19	2.389
8	117	690	030_11:32	711_15:13	1299.09	1.883
9	131	1015	028_14:41	699_19:11	3205.24	3.158
10	152	223	038_18:02	696_19:11	755.46	3.388

- 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)

```

select
d.*
from `ecommerce.transaction` t join `ecommerce.demographics` d on
t.household_key = d.household_key
order by t.SALES_VALUE desc
limit 1

```

Row	AGE_DESC	MARITAL_STATUS	INCOME_DESC	HOMEOWNER_DESC	HH_COMP_DESC	HOUSEHOLD_SIZE_DESC	KID_CATEGORY_DESC	household_key
1	45-54	A	125-149K	Homeowner	2 Adults Kids	5+	3+	1609

- Insights : Most spending customer is from 45 to 54 age group. We can say that these people are financial independent. Their Income group is 125 to 149k which is very high. They are having 3+ kids and family size is 5+. Household key is 1609.
- **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)

```

WITH cte1 AS (
    SELECT
        t.product_id,
        t.BASKET_ID,
        p.SUB_COMMODITY_DESC
    FROM
        `ecommerce.transaction` t
        JOIN `ecommerce.product` p USING(product_id)
),
cte2 AS (
    SELECT
        DISTINCT
        a.BASKET_ID,
        CASE
            WHEN a.SUB_COMMODITY_DESC < b.SUB_COMMODITY_DESC
            THEN a.SUB_COMMODITY_DESC
            ELSE b.SUB_COMMODITY_DESC
        END AS product1,
        CASE
            WHEN a.SUB_COMMODITY_DESC < b.SUB_COMMODITY_DESC
            THEN b.SUB_COMMODITY_DESC
            ELSE a.SUB_COMMODITY_DESC
        END AS product2
    FROM
        cte1 a
        JOIN cte1 b
        ON a.BASKET_ID = b.BASKET_ID
WHERE

```

```

        a.SUB_COMMODITY_DESC <> b.SUB_COMMODITY_DESC
    )
SELECT
    product1,
    product2,
    COUNT(*) AS combination_count
FROM
    cte2
GROUP BY
    product1, product2
ORDER BY
    combination_count DESC;

```

Row	product1	product2	combination_count
1	BANANAS	FLUID MILK WHITE ONLY	4131
2	FLUID MILK WHITE ONLY	MAINSTREAM WHITE BREAD	3753
3	FLUID MILK WHITE ONLY	SOFT DRINKS 12/18&15PK CA...	3328
4	FLUID MILK WHITE ONLY	SHREDDED CHEESE	3155
5	FLUID MILK WHITE ONLY	YOGURT NOT MULTI-PACKS	2805
6	DAIRY CASE 100% PURE JUICE...	FLUID MILK WHITE ONLY	2682
7	FLUID MILK WHITE ONLY	SFT DRNK 2 LITER BTL CARB I...	2579
8	FLUID MILK WHITE ONLY	KIDS CEREAL	2554
9	FLUID MILK WHITE ONLY	POTATO CHIPS	2200
10	EGGS - LARGE	FLUID MILK WHITE ONLY	1952
11	BANANAS	YOGURT NOT MULTI-PACKS	1931

Insights : Above is the pair wise combination for the products which brought together. The most coupled products are Bananas and Fluid Milk only.

- 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
    d.household_key AS customer_key, t.WEEK_NO,
    round(sum(t.SALES_VALUE),2) AS current_week_value,
from `ecommerce.demographics` d
JOIN `ecommerce.transaction` t
ON d.household_key = t.household_key
group by 1,2)

select

```

```
cte.customer_key AS CUSTOMER_KEY, WEEK_NO, cte.current_week_value
AS CURRENT_WEEK_VALUE,
LAG(cte.current_week_value) OVER (PARTITION BY cte.customer_key
ORDER BY WEEK_NO) AS PREV_WEEK_VALUE,
round((cte.current_week_value) - (LAG(cte.current_week_value) OVER
(PARTITION BY cte.customer_key ORDER BY WEEK_NO)),2) AS
WEEKLY_DIFFERENCE
FROM cte
```

Row	CUSTOMER_KEY	WEEK_NO	CURRENT_WEEK_VA	PREV_WEEK_VALUE	WEEKLY_DIFFERENC
1	216	1	10.39	null	null
2	216	9	3.67	10.39	-6.72
3	216	11	15.47	3.67	11.8
4	216	12	7.49	15.47	-7.98
5	216	13	3.12	7.49	-4.37
6	216	16	23.85	3.12	20.73
7	216	17	2.79	23.85	-21.06
8	216	24	2.99	2.79	0.2
9	216	25	4.82	2.99	1.83
10	216	26	8.79	4.82	3.97
11	216	28	22.64	8.79	13.85

Extra Questions

Question – 1 : Find the total sale in descending order by age category.

```
select
AGE_DESC as Age_Group,
round(sum(SALES_VALUE),2) as Total_Sale
from `ecommerce.demographics` d join `ecommerce.transaction` t on
d.household_key = t.household_key
group by 1
order by 2 desc
```

Row	Age_Group	Total_Sale
1	45-54	827984.9
2	35-44	622164.35
3	25-34	389545.17
4	65+	151606.81
5	55-64	150371.27
6	19-24	108404.35

Insights : Maximum sale is done by the age_group 45 – 54 so, we can say that this age group person are financially independent. Also, we observed that between 19 to 24 age_group, purchase is least, as this age group person and not independent and rely on parents for expenses.

Question – 2 : Find out the total revenue by Marital Status and different Income Groups.

```
select
INCOME_DESC as Income_Group, MARITAL_STATUS_CODE,
round(sum(SALES_VALUE),2) as Total_Sale
from `ecommerce.demographics` d join `ecommerce.transaction` t on
d.household_key = t.household_key
group by 1,2
order by 1, 3
```

Row	Income_Group ▼	MARITAL_STATUS_CODE ▼	Total_Sale ▼
1	100-124K	B	12144.2
2	100-124K	U	19395.09
3	100-124K	A	69392.18
4	125-149K	B	10109.64
5	125-149K	U	39203.64
6	125-149K	A	101151.42
7	15-24K	B	17232.45
8	15-24K	A	42659.57
9	15-24K	U	91448.26
10	150-174K	B	13780.72

Insights : In every income_group, maximum sale is done by the married person only.

Question – 3: Find out total sale as per the house owner description.

```
select
HOMEOWNER_DESC,
round(sum(SALES_VALUE),2) as Total_Sale
from `ecommerce.demographics` d join `ecommerce.transaction` t on
d.household_key = t.household_key
group by 1
order by 2 desc
```

Row	HOMEOWNER_DESC	Total_Sale
1	Homeowner	1519166.59
2	Unknown	561266.78
3	Renter	118735.77
4	Probable Owner	27236.49
5	Probable Renter	23671.22

Insights : Most of the sale is done by homeowner and least sale is by probable renter.

Question – 4: Find out Total Sale by every manufacturer with their total quantity.

```
select
MANUFACTURER as Manufacturer, count(p.PRODUCT_ID) as Total_qty,
round(sum(SALES_VALUE),2) as Total_Sale
from `ecommerce.product` p join `ecommerce.transaction` t on
p.PRODUCT_ID = t.PRODUCT_ID
group by 1
order by 2 desc, 2
```

Row	Manufacturer	Total_qty	Total_Sale
1	69	366825	1090508.12
2	2	85429	174160.8
3	544	23249	51087.33
4	317	19652	51754.18
5	103	18989	59227.85
6	1251	17716	42872.18
7	1208	17379	57409.78
8	1046	17221	35361.51
9	764	15647	83073.14
10	693	14267	20775.18
11	673	13098	41710.46

Question – 5: Find out department wise total sale.

```
select
DEPARTMENT as Department,
round(sum(SALES_VALUE),2) as Total_Sale
from `ecommerce.product` p join `ecommerce.transaction` t on
p.PRODUCT_ID = t.PRODUCT_ID
group by 1
order by 2 desc
```

Row	Department	Total_Sale
1	GROCERY	2046695.13
2	DRUG GM	527588.65
3	PRODUCE	279720.39
4	MEAT	274036.32
5	KIOSK-GAS	269461.67
6	MEAT-PCKGD	206491.71
7	DELI	130322.26
8	MISC SALES TRAN	62633.95
9	PASTRY	61786.56
10	NUTRITION	48840.39

Insights : Most of the sale is done by Grocery department and least is by nutrition.

Question – 6: Calculate the total revenue generated per household and categorize it by income level, marital status, and age group. Identify the top 10 households contributing the most revenue in each income category.

```
with cte as
(select
d.household_key, INCOME_DESC, MARITAL_STATUS_CODE, AGE_DESC,
round(sum(t.SALES_VALUE),2) AS total_revenue
from `ecommerce.demographics` d
JOIN `ecommerce.transaction` t
ON d.household_key = t.household_key
group by 1,2,3,4)
```

```

select household_key AS HOUSEHOLD_KEY, INCOME_DESC, MARITAL_STATUS_CODE,
AGE_DESC, total_revenue as TOTAL_REVENUE from
(select
household_key, INCOME_DESC, MARITAL_STATUS_CODE, AGE_DESC,
total_revenue,
rank() over (Partition by INCOME_DESC order by total_revenue desc) as
rnk
from cte) aa
where rnk <= 10
order by total_revenue DESC

```

Row	HOUSEHOLD_KEY	INCOME_DESC	MARITAL_STATUS_CODE	AGE_DESC	TOTAL_REVENUE
1	1609	125-149K	A	45-54	13804.38
2	2322	175-199K	U	45-54	11934.66
3	1453	125-149K	A	45-54	10720.72
4	1430	35-49K	A	35-44	10147.21
5	718	25-34K	A	45-54	9577.63
6	1653	Under 15K	B	35-44	9519.93
7	400	150-174K	A	35-44	9481.19
8	982	35-49K	U	45-54	9388.07
9	707	100-124K	A	25-34	9364.74
10	1229	150-174K	A	55-64	9256.85
11	1527	50-74K	A	25-34	8864.4

Question – 7 : For each product department, calculate the total sales, average product price, and the number of unique products sold. Determine which department has the highest revenue and which has the highest average price.

```

with cte as
(select DEPARTMENT,
sum(SALES_VALUE) as Revenue, count(distinct p.PRODUCT_ID) as
product_count, avg(SALES_VALUE) as Avg_price
from `ecommerce.transaction` t join `ecommerce.product` p on
t.PRODUCT_ID = p.PRODUCT_ID
group by 1)

select
max(case when Revenue = (select max(cte.Revenue) from cte) then
Department end) as Highest_Revenue_dept,
max(case when cte.Avg_price = (select max(cte.Avg_price) from cte) then
Department end) as Highest_avg_sales_dept
from cte

```

Row	Highest_Revenue_dept	Highest_avg_sales_dept
1	GROCERY	KIOSK-GAS

Insights : Highest revenue is generated by Grocery and highest avg sales dept is Kiosk-Gas

Question – 8 : Determine how household characteristics (such as the number of children and homeownership) affect the total revenue generated. Identify any significant trends or correlations.

```
select
HOMEOWNER_DESC, HH_COMP_DESC,
round(sum(SALES_VALUE),2) as Total_Sale
from `ecommerce.demographics` d join `ecommerce.transaction` t on
d.household_key = t.household_key
group by 1, 2
order by 2 desc
```

Row	HOMEOWNER_DESC ▾	HH_COMP_DESC ▾	Total_Sale ▾
1	Homeowner	Unknown	103967.73
2	Unknown	Unknown	75354.38
3	Probable Renter	Unknown	6290.78
4	Renter	Unknown	5659.21
5	Unknown	Single Male	105122.62
6	Homeowner	Single Male	100851.46
7	Renter	Single Male	19077.4
8	Probable Renter	Single Male	3744.09
9	Probable Renter	Single Female	13636.35
10	Unknown	Single Female	170599.65

Question – 9: For each income group, identify the top 10 most popular products based on quantity sold. Calculate the percentage contribution of these products to the group's total revenue.

```
with cte1 as
(select income_desc, sum(quantity) as total from
`ecommerce.demographics` d join `ecommerce.transaction` t on
d.household_key = t.household_key group by 1),
cte as
(select
INCOME_DESC, PRODUCT_ID,
sum(QUANTITY) as TOTAL_QUANTITY
from `ecommerce.demographics` d join `ecommerce.transaction` t on
d.household_key = t.household_key
group by 1, 2)

select
income_desc, product_id, TOTAL_QUANTITY, percentage_contribution
from (select
income_desc, product_id, cte.TOTAL_QUANTITY,
dense_rank() over (partition by income_desc order by cte.TOTAL_QUANTITY
desc) as rnk,
```

```

round(100*(total_quantity / (select total from cte1 where
cte1.income_desc = cte.income_desc)),2 ) as percentage_contribution from
cte ) aa
where rnk <= 10
order by income_desc

```

Row	income_desc	product_id	TOTAL_QUANTITY	percentage_contribution
1	100-124K	6534178	4039726	75.6
2	100-124K	6533889	1110011	20.77
3	100-124K	6544236	115431	2.16
4	100-124K	397896	12700	0.24
5	100-124K	5703832	11840	0.22
6	100-124K	6534166	9381	0.18
7	100-124K	1404121	6805	0.13
8	100-124K	1082185	508	0.01
9	100-124K	995242	227	0.0
10	100-124K	5587133	211	0.0
11	125-149K	6534178	4828804	72.92

Question – 10: Analyze how coupon usage differs across households of different sizes. Determine if larger households are more likely to use coupons and if this leads to higher overall savings or spending.

```

SELECT
    HOUSEHOLD_SIZE_DESC AS household_size,
    ROUND(SUM(COUPON_DISC), 2) AS total_coupon_savings,
    ROUND(SUM(SALES_VALUE), 2) AS total_spending,
    COUNT(t.BASKET_ID) AS total_transactions,
    ROUND(SUM(COUPON_DISC) / SUM(SALES_VALUE) * 100, 2) AS
coupon_savings_percentage
FROM `ecommerce.demographics` d JOIN `ecommerce.transaction` t
ON d.household_key = t.household_key
GROUP BY 1
ORDER BY total_coupon_savings DESC;

```

Row	household_size	total_coupon_saving	total_spending	total_transactions	coupon_savings_percentage
1	4	-1373.75	175637.09	56498	-0.78
2	5+	-1603.51	221412.68	67065	-0.72
3	3	-2067.89	332013.32	106530	-0.62
4	1	-3601.29	640187.6	210533	-0.56
5	2	-5119.97	880826.16	273231	-0.58