Analysis of Super-Market using SQL

- **Finding from Super-Market Dataset**

  Super-Market Effectively managed operations in Brazil, overseeing the processing of approximately 100,000 orders spanning from September 2016 to October 2018.

  - Derived insights that customers orders from 4,119 cities across 27 states during this period, with São Paulo (SP) accounting for 40% of total customer base.
  - Analyzed that avg 19% growth in order volume and a significant 143% increase in order cost from 2017 to 2018.
  - Identified peak order months (Jan-Aug) constituting 76% of total orders, with 66% placed during afternoon and night and only 5% during dawn
  - Analyzed that company can improve delivery time of product , some product reach after estimated delivery time.
  - Analyzed payment options, revealing that 82% of orders offered less than 3 EMI options, suggesting potential improvements in payment flexibility.

QUES 1. Import the dataset and do usual exploratory analysis steps like checking the structure & characteristics of the dataset:

ANS 1. (A) Data type of all columns in the "customers" table.

Assumption: used information schema to get the data type of all column and used where clause to get the data type of customer table

ANS.
```
SELECT column_name,data_type
FROM target-410104.Target.INFORMATION_SCHEMA.COLUMNS
where table_name ="customers"
```

Row	column_name	data_type
1	customer_id	STRING
2	customer_unique_id	STRING
3	customer_zip_code_prefix	INT64
4	customer_city	STRING
5	customer_state	STRING

(B) Get the time range between which the orders were placed.

Assumption: using order table to get the first order and last order of the given datasets

With the help of aggregate function (min,max) and present the first and last order in timestamp format.

ANS.
```
select min(order_purchase_timestamp) as First_order,
max(order_purchase_timestamp) as Last_order
from `target-410104.Target.orders`
```

Row	First_order	Last_order
1	2016-09-04 21:15:19 UTC	2018-10-17 17:30:18 UTC

(C) Count the Cities & States of customers who ordered during the given period.
Assumption: using order and customer table and join them using inner join because I want only those cities and states of customers who ordered during this time , after joining them use count and distinct function to count unique cities and states

```
ANS . SELECT COUNT(DISTINCT(c.customer_city)) AS City_Count,
      COUNT(DISTINCT(c.customer_state)) AS State_Count
      FROM `target-410104.Target.customers` AS c
      JOIN `target-410104.Target.orders` AS o
      ON o.customer_id = c.customer_id
```

Row	City_Count	State_Count
1	4119	27

QUES 2. In-depth Exploration:

Ans 2 (A) Is there a growing trend in the no. of orders placed over the past years?
Assumption: using order table and group the year and count the distinct order to get the number of order placed in those years

```
ANS. select extract(Year from order_purchase_timestamp) as years,
     count(distinct order_id) as Number_of_Orders
     from `target-410104.Target.orders`
     group by years
     order by years
```

Row	years	Number_of_Orders
1	2016	329
2	2017	45101
3	2018	54011

Based on the analysis, if there's a clear increased in the number of orders placed over the past years, the recommendation would be to:

- **Scale Operations**: Prepare for increased demand by scaling up operational capacities, ensuring sufficient inventory levels, and optimizing logistics and delivery systems to meet growing customer demands.
- **Marketing Strategies**: Leverage this growth trend by investing in targeted marketing campaigns, promotions, or loyalty programs to retain existing customers and attract new ones.
- **Customer Service Enhancement**: Focus on improving customer service and experience to accommodate the increasing number of orders efficiently.

(B) Can we see some kind of monthly seasonality in terms of the no. of orders being placed?

Assumption: group by months and gets the total order placed in specific months of combining years to know which get the most orders but this data is from September 2016 to October 2018 so might be so fluctuation on number of orders at aggregate months level.

```
select  extract(Month from order_purchase_timestamp) as months,
 count(distinct order_id) as No_of_Orders
 from `target-410104.Target.orders`
 group by months
 order by months
```

Row	months	No_of_Orders
1	1	8069
2	2	8508
3	3	9893
4	4	9343
5	5	10573
6	6	9412
7	7	10318
8	8	10843
9	9	4305
10	10	4959
11	11	7544
12	12	5674

There's a noticeable trend of increasing orders from March to August, it suggests a seasonal surge in demand during these months. Some recommendations based on this observation:

- Anticipate and prepare for increased demand by adjusting inventory levels for popular products. Ensure adequate stock availability to meet customer needs during these months
- During peak months to handle higher order volumes efficiently. Consider scaling up workforce or optimizing operational processes to manage increased demand without compromising service quality.
- Align marketing campaigns and promotional activities to leverage this seasonal trend.
- Enhance customer engagement initiatives during these months. Offer personalized recommendations, loyalty programs, or special offers to encourage for purchases .

Assumption: group data by years and month level and get the number of orders placed in one specific month year wise , to get the year and month from timestamp used the extract function

ANS .
```
select extract(Year from order_purchase_timestamp) as years,
 extract(Month from order_purchase_timestamp) as months,
 count(distinct order_id) as Number_of_Orders
 from `target-410104.Target.orders`
 group by years, months
 order by years, months
```

Row	years	months	No_of_Orders
1	2016	9	4
2	2016	10	324
3	2016	12	1
4	2017	1	800
5	2017	2	1780
6	2017	3	2682
7	2017	4	2404
8	2017	5	3700
9	2017	6	3245
10	2017	7	4026
11	2017	8	4331
12	2017	9	4285

Row	years	months	No_of_Orders
13	2017	10	4631
14	2017	11	7544
15	2017	12	5673
16	2018	1	7269
17	2018	2	6728
18	2018	3	7211
19	2018	4	6939
20	2018	5	6873
21	2018	6	6167
22	2018	7	6292
23	2018	8	6512
24	2018	9	16
25	2018	10	4

(C). During what time of the day, do the Brazilian customers mostly place their orders? (Dawn, Morning, Afternoon or Night)

- 0-6 hrs : Dawn
- 7-12 hrs : Mornings
- 13-18 hrs : Afternoon
- 19-23 hrs : Night

ANS .
```
select
  sum(case when Hours between 0 and 6 then Total_orders else 0 end) as Dawn,
  sum(case when Hours between 7 and 12 then Total_orders else 0 end) as Mornings,
  sum(case when Hours between 13 and 18 then Total_orders else 0 end) as Afternoon,
  sum(case when Hours between 19 and 23 then Total_orders else 0 end) as Night
from( select extract(hour from order_purchase_timestamp) as Hours,
    count(distinct order_id) as Total_orders
    from `target-410104.Target.orders`
    group by Hours
    order by Hours) as t
```

Row	Dawn	Mornings	Afternoon	Night
1	5242	27733	38135	28331

- Based on analysis it indicates a maximum orders during the afternoon, its advisable to concentrate marketing efforts or promotional campaigns during that time to capture maximum customer engagement. Similarly, ensuring optimal staffing and logistics support during peak ordering hours can enhance customer satisfaction by expediting order processing and delivery.

QUES.3. Evolution of E-commerce orders in the Brazil region:

Ans 3. (A) Get the month on month no. of orders placed in each state.

ANS.
```
select c.customer_state as State,
  extract(Year from order_purchase_timestamp) as years,
  extract(Month from o.order_purchase_timestamp) as months,
  count(distinct o.order_id) as No_of_Orders
  from `target-410104.Target.customers` as c
```

```
join `target-410104.Target.orders` as o
on c.customer_id = o.customer_id
group by c.customer_state,years,months
order by State,years,months
```

Row	State	years	months	No_of_Orders
1	AC	2017	1	2
2	AC	2017	2	3
3	AC	2017	3	2
4	AC	2017	4	5
5	AC	2017	5	8
6	AC	2017	6	4
7	AC	2017	7	5
8	AC	2017	8	4
9	AC	2017	9	5
10	AC	2017	10	6

Some states exhibit consistent growth,or some show irregular patterns in order placements

(A) How are the customers distributed across all the states?

ANS.
```
select customer_state as State,
count(distinct customer_unique_id) as Total_customers
from `target-410104.Target.customers`
group by customer_state
order by Total_customers desc
```

Row	State	Total_customers
1	SP	40302
2	RJ	12384
3	MG	11259
4	RS	5277
5	PR	4882
6	SC	3534
7	BA	3277
8	DF	2075
9	ES	1964
10	GO	1952

By analyzing and understanding customer distribution across states, Target can tailor its strategies, operations, and marketing efforts to top states market and optimize its presence in Brazil.

QUES.4 Impact on Economy: Analyze the money movement by e-commerce by looking at order prices, freight and others.

Ans .4

(A) Get the % increase in the cost of orders from year 2017 to 2018 (include months between Jan to Aug only).You can use the "payment_value" column in the payments table to get the cost of orders.

ANS.
```
WITH Refernce_table AS (
  SELECT EXTRACT(YEAR FROM order_purchase_timestamp) AS years,
```

```
round(SUM(payment_value),0) AS cost
FROM `target-410104.Target.orders` AS o
JOIN `target-410104.Target.payments` AS p ON o.order_id = p.order_id
WHERE EXTRACT(YEAR FROM o.order_purchase_timestamp) BETWEEN 2017 AND 2018
 AND EXTRACT(MONTH FROM o.order_purchase_timestamp) BETWEEN 1 AND 8
 and o.order_status="delivered"
 GROUP BY years),

Final_table AS (
 SELECT *,
 LAG(cost) OVER (ORDER BY years) as prev_year_cost
 FROM Refernce_table)

SELECT *,
round((cost - prev_year_cost) * 100 / prev_year_cost,0) AS perce_increase
FROM Final_table;
```

Row	years	cost	prev_year_cost	perce_increase
1	2018	8452975.0	3473863.0	143.0
2	2017	3473863.0	*null*	*null*

Based on the analysis there is increase in cost of order from last year so its a profitable place for Target to do business , Target can strategize effectively to sustain or improve this growth trajectory in the future.

(B) Calculate the Total & Average value of order price for each state.

ANS.
```
select c.customer_state as State,
 round(sum(oi.price),0) as total_value_of_price,
 round(avg(oi.price),0) as avg_value_of_price
 from `target-410104.Target.customers` as c
 join `target-410104.Target.orders` as o
 on c.customer_id = o.customer_id
 join `target-410104.Target.order_items` as oi
 on o.order_id = oi.order_id
 where o.order_status="delivered"
 group by c.customer_state
 order by avg_value_of_price desc,total_value_of_price desc
```

Row	State	total_value_of_price	avg_value_of_price
1	PB	112587.0	192.0
2	AL	78856.0	185.0
3	AC	15931.0	175.0
4	RO	45683.0	167.0
5	PA	174471.0	166.0
6	AP	13375.0	165.0
7	PI	84721.0	162.0
8	RN	82106.0	158.0
9	TO	48403.0	156.0
10	CE	219757.0	154.0

Target understand customer behavior and preferences in these high-spending states through further analysis. Tailor marketing efforts, product assortments, and customer

service strategies to meet the specific demands and expectations of customers in these regions, ultimately driving higher sales and customer satisfaction.

(C) Calculate the Total & Average value of order freight for each state.

```
ANS. select c.customer_state as State,
    round(sum(oi.freight_value),0) as total_freight_value,
    round(avg(oi.freight_value),0) as avg_freight_value
    from `target-410104.Target.customers` as c
    join `target-410104.Target.orders` as o
    on c.customer_id = o.customer_id
    join `target-410104.Target.order_items` as oi
    on o.order_id = oi.order_id
    where o.order_status = "delivered"
    group by c.customer_state
    order by avg_freight_value desc ,total_freight_value desc
```

Row	State	total_freight_value	avg_freight_value
1	PB	25252.0	43.0
2	RR	1982.0	43.0
3	RO	11283.0	41.0
4	AC	3644.0	40.0
5	PI	20457.0	39.0
6	MA	30794.0	38.0
7	SE	13715.0	37.0
8	TO	11605.0	37.0
9	PA	37553.0	36.0
10	RN	18609.0	36.0

For states with the highest average freight values, investigate the underlying reasons such as distance from distribution centers, transportation infrastructure, or carrier preferences. Target might consider optimizing logistics routes, negotiating better rates with carriers, or establishing regional distribution centers to reduce costs.

QUES . 5. **Analysis based on sales, freight and delivery time.**

Ans.5

(A) Find the no. of days taken to deliver each order from the order's purchase date as delivery time. Also, calculate the difference (in days) between the estimated & actual delivery date of an order.Do this in a single query.
You can calculate the delivery time and the difference between the estimated & actual delivery date using the given formula:

i. **time_to_deliver** = order_delivered_customer_date - order_purchase_timestamp
ii. **diff_estimated_delivery** = order_estimated_delivery_date - order_delivered_customer_date

```
ANS. select order_id,
    timestamp_diff(order_delivered_customer_date,order_purchase_timestamp,day) as time_to_delivery,
    timestamp_diff(order_estimated_delivery_date,order_delivered_customer_date,day) as
    diff_estimated_delivery
    from `target-410104.Target.orders`
```

```
where order_status="delivered"
order by time_to_delivery desc
```

Row	order_id	time_to_delivery	diff_estimated_delivery
1	ca07593549f1816d26a572e06dc1eab6	209	-181
2	1b3190b2dfa9d789e1f14c05b647a14a	208	-188
3	440d0d17af552815d15a9e41abe49359	195	-165
4	0f4519c5f1c541ddec9f21b3bddd533a	194	-161
5	285ab9426d6982034523a855f55a885e	194	-166
6	2fb597c2f772eca01b1f5c561bf6cc7b	194	-155
7	47b40429ed8cce3aee9199792275433f	191	-175
8	2fe324febf907e3ea3f2aa9650869fa5	189	-167
9	2d7561026d542c8dbd8f0daeadf67a43	188	-159
10	437222e3fd1b07396f1d9ba8c15fba59	187	-144

- Identify bottlenecks, inefficiencies, or delays in processing orders, preparing shipments, or transportation. Streamlining these processes can significantly reduce delivery times.
- Ensure sufficient stock availability and accurate inventory levels to prevent delays caused by stockouts or backorders. .
- Proactively communicate with customers about potential delays or changes in delivery schedules. Providing accurate and timely updates can manage expectations and enhance customer satisfaction despite delays.
- Implement robust monitoring systems to track delivery performance continually

(B) Find out the top 5 states with the highest & lowest average freight value

ANS.
```
select c.customer_state as State,
round(avg(oi.freight_value),0) as avg_freight_value
from `target-410104.Target.customers` as c
join `target-410104.Target.orders` as o
on c.customer_id = o.customer_id
join `target-410104.Target.order_items` as oi
on o.order_id = oi.order_id
group by c.customer_state
order by avg_freight_value desc
limit 5
```

Row	State	avg_freight_value
1	PB	43.0
2	RR	43.0
3	RO	41.0
4	AC	40.0
5	PI	39.0

```
select c.customer_state as State,
round(avg(oi.freight_value),0) as avg_freight_value
from `target-410104.Target.customers` as c
join `target-410104.Target.orders` as o
```

```
on c.customer_id = o.customer_id
join `target-410104.Target.order_items` as oi
on o.order_id = oi.order_id
group by c.customer_state
order by avg_freight_value
limit 5
```

Row	State	avg_freight_value
1	SP	15.0
2	PR	21.0
3	SC	21.0
4	RJ	21.0
5	MG	21.0

(C) Find out the top 5 states with the highest & lowest average delivery time.

ANS.
```
select distinct c.customer_state,
    round(avg(o.time_to_delivery) over(partition by c.customer_state),0)  as time_to_delivery
from `target-410104.Target.customers` as c
join (select *,
      timestamp_diff(order_delivered_customer_date,order_purchase_timestamp,day) as
    time_to_delivery
      from `target-410104.Target.orders`
      where order_status="delivered"
      order by time_to_delivery) as o
on c.customer_id = o.customer_id
order by time_to_delivery desc
limit 5
```

Row	customer_state	time_to_delivery
1	RR	29.0
2	AP	27.0
3	AM	26.0
4	AL	24.0
5	PA	23.0

```
select distinct c.customer_state,round(avg(o.time_to_delivery) over(partition by c.customer_state),0) as
time_to_delivery
from `target-410104.Target.customers` as c
join (select *,
timestamp_diff(order_delivered_customer_date,order_purchase_timestamp,day) as time_to_delivery
from `target-410104.Target.orders`
where order_status="delivered"
order by time_to_delivery) as o
on c.customer_id = o.customer_id
order by time_to_delivery
limit 5
```

Row	customer_state	time_to_delivery
1	SP	8.0
2	MG	12.0
3	PR	12.0
4	DF	13.0
5	SC	14.0

(D) Find out the top 5 states where the order delivery is really fast as compared to the estimated date of delivery.
You can use the difference between the averages of actual & estimated delivery date to figure out how fast the delivery was for each state.

```
ANS. select distinct c.customer_state,
     round(avg(o.diff_estimated_delivery) over(partition by c.customer_state),0) as diff_estimated_delivery
     from `target-410104.Target.customers` as c
     join (select *,
          timestamp_diff(order_estimated_delivery_date,order_delivered_customer_date,day) as
            diff_estimated_delivery
          from `target-410104.Target.orders`
          where order_status="delivered"
          order by diff_estimated_delivery)as o
     on c.customer_id = o.customer_id
     order by diff_estimated_delivery
     limit 5
```

Row	customer_state	diff_estimated_delivery
1	AL	8.0
2	SE	9.0
3	MA	9.0
4	BA	10.0
5	SP	10.0

QUES 6. Analysis based on the payments:

Ans.6

(A) Find the month on month no. of orders placed using different payment types.

```
ANS.  with cte as
          (select *,
          lag(t.No_of_Orders) over(partition by t.Payment_type order by t.years , t.months) as prev_order
          from(select p.payment_type as Payment_type,
                    extract(Year from o.order_purchase_timestamp) as years,
                    extract(Month from o.order_purchase_timestamp) as months,
                    count(distinct o.order_id) as No_of_Orders
                    from `target-410104.Target.payments` as p
                    join `target-410104.Target.orders` as o
                    on p.order_id = o.order_id
                    group by years,months,p.payment_type
                    order by Payment_type,years,months) as t
          order by Payment_type,years,months asc),
```

```
cte2 as(select *,
(No_of_Orders - prev_order) as diff_in_order
from cte)

select *,
round((diff_in_order/prev_order) * 100,0) as perc_diff_in_order
from cte2
```

Row	Payment_type	years	months	No_of_Orders	prev_order	diff_in_order	perc_diff_in_order
1	UPI	2016	10	63	null	null	null
2	UPI	2017	1	197	63	134	213.0
3	UPI	2017	2	398	197	201	102.0
4	UPI	2017	3	590	398	192	48.0
5	UPI	2017	4	496	590	-94	-16.0
6	UPI	2017	5	772	496	276	56.0
7	UPI	2017	6	707	772	-65	-8.0
8	UPI	2017	7	845	707	138	20.0
9	UPI	2017	8	938	845	93	11.0
10	UPI	2017	9	903	938	-35	-4.0

Popularity of different payment types among customers. If certain types are less utilized, consider adding or promoting alternative payment methods that are more widely used or preferred by customers.

(B) Find the no. of orders placed on the basis of the payment installments that have been paid.

ANS.
```
select payment_installments,count(order_id) as no_of_orders
from`target-410104.Target.payments`
where payment_installments >1
group by payment_installments
order by no_of_orders desc
```

Row	payment_installments	no_of_orders
1	2	12413
2	3	10461
3	4	7098
4	10	5328
5	5	5239
6	8	4268
7	6	3920
8	7	1626
9	9	644
10	12	133
11	15	74
12	18	27
13	11	23
14	24	18

Row	payment_installments	no_of_orders
14	24	18
15	20	17
16	13	16
17	14	15
18	17	8
19	16	5
20	21	3
21	22	1
22	23	1

Based on analysis number of orders is usually high when there is less installments.