1. Import the dataset and do usual exploratory analysis steps like checking the structure & characteristics of the dataset
    1. Data type of columns in a table

```
SELECT COLUMN_NAME,DATA_TYPE

FROM target.INFORMATION_SCHEMA.COLUMNS
where table_name='orders'
```

Row	COLUMN_NAME	DATA_TYPE
1	order_id	STRING
2	customer_id	STRING
3	order_status	STRING
4	order_purchase_timestamp	TIMESTAMP
5	order_approved_at	TIMESTAMP
6	order_delivered_carrier_date	TIMESTAMP
7	order_delivered_customer_date	TIMESTAMP
8	order_estimated_delivery_date	TIMESTAMP

    2. Time period for which the data is given

```
select min(order_purchase_timestamp) as start_time,
max(order_purchase_timestamp) as end_time from target.orders
```

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

    3. Cities and States of customers ordered during the given period

```
select c.customer_city,c.customer_state from target.customers as c
join target.orders as o on c.customer_id=o.customer_id
group by 2,1 limit 10
```

Row	customer_city	customer_state
1	acu	RN
2	ico	CE
3	ipe	RS
4	ipu	CE
5	ita	SC
6	itu	SP
7	jau	SP
8	luz	MG
9	poa	SP
10	uba	MG

2. In-depth Exploration:
    1. Is there a growing trend on e-commerce in Brazil? How can we describe a complete scenario? Can we see some seasonality with peaks at specific months?

```
SELECT extract(year from order_purchase_timestamp) as Year,
extract(month from order_purchase_timestamp) as Month,
count(distinct(order_id)) as Total_orders
FROM target.orders group by 1,2 order by 1,2
```

Row	Year	Month	Total_orders
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
13	2017	10	4631
14	2017	11	7544

2. What time do Brazilian customers tend to buy (Dawn, Morning, Afternoon or Night)?

```
select sum(CASE when hours between 00 and 06 then orders end)as
Dawn,
sum(CASE when hours between 07 and 12 then orders end)as Morning,
sum(CASE when hours between 13 and 19 then orders end)asAfternoon,
sum(CASE when hours between 20 and 23 then orders end)as Night
from (
SELECT extract(Hour from order_purchase_timestamp) as hours,
count(distinct(order_id)) as orders FROM target.orders
group by 1 )as x
```

Row	Dawn	Morning	Afternoon	Night
1	5242	27733	44117	22349

3. Evolution of E-commerce orders in the Brazil region:
    1. Get month on month orders by states

```
select extract(Month from o.order_purchase_timestamp) as Months,
count(distinct(o.order_id)) as Total_orders,
c.customer_state from target.customers as c
join target.orders as o
on c.customer_id=o.customer_id
group by 1,c.customer_state
order by 1,2
```

Row	Months	Total_orders	customer_state
1	1	2	RR
2	1	8	AC
3	1	11	AP
4	1	12	AM
5	1	19	TO
6	1	23	RO
7	1	24	SE
8	1	33	PB
9	1	39	AL
10	1	51	RN

2. Distribution of customers across the states in Brazil

```
select customer_state,customer_city,
count(distinct(customer_unique_id)) as Total_customers
from target.customers
group by customer_state,customer_city
order by 3 desc
```

Row	customer_state	customer_city	Total_customers
1	SP	sao paulo	14984
2	RJ	rio de janeiro	6620
3	MG	belo horizonte	2672
4	DF	brasilia	2069
5	PR	curitiba	1465
6	SP	campinas	1398
7	RS	porto alegre	1326
8	BA	salvador	1209
9	SP	guarulhos	1153
10	SP	sao bernardo do campo	908

4. Impact on Economy: Analyze the money movement by e-commerce by looking at order prices, freight and others.
    1. Get % increase in cost of orders from 2017 to 2018 (include months between Jan to Aug only) - You can use “payment_value” column in payments table

```
with base as (
select * from target.orders a
join target.payments b on a.order_id = b.order_id
where extract(year from a.order_purchase_timestamp) between 2017 and 2018
and extract(month from a.order_purchase_timestamp) between 1 and 8),
base2 as (
select extract(year from order_purchase_timestamp) as year, sum(payment_value)
as cost from base
group by 1 order by 1 asc),
base3 as (
select *, lead(cost, 1) over (order by year) as next_year_cost from base_2)
select *, (next_year_cost - cost)/ cost *100 as per_inc from base_3
```

Row	year	cost	next_year_cost	per_inc
1	2018	8694733.83...	*null*	*null*
2	2017	3669022.11...	8694733.83...	136.976871...

2. Mean & Sum of price and freight value by customer state

```
SELECT sum(oi.price) as Sum_price,Avg(oi.price)avg_price,c.customer_state,
oi.freight_value FROM target.customers c
join target.orders o
on c.customer_id=o.customer_id
join target.order_items oi
on o.order_id=oi.order_id
group by c.customer_state,oi.freight_value
order by 4 desc
```

Row	Sum_price	avg_price	customer_state	freight_value
1	979.0	979.0	PI	409.68
2	2338.08	2338.08	PR	375.28
3	2338.08	2338.08	SC	375.28
4	1149.0	1149.0	SP	339.59
5	1050.0	1050.0	MT	338.3
6	1050.0	1050.0	MG	322.1
7	1050.0	1050.0	ES	321.88
8	990.0	990.0	SP	321.46
9	3089.0	3089.0	PB	317.47
10	1045.0	1045.0	AL	314.4

5. Analysis on sales, freight and delivery time

1. Calculate days between purchasing, delivering and estimated delivery

```
select date_diff(o.order_purchase_timestamp,o.order_delivered_customer_date,
day) as time_to_delivery,
date_diff(o.order_estimated_delivery_date,o.order_purchase_timestamp,day) as
diff_estimated_delivery from target.customers as c
join target.orders as o on c.customer_id=o.customer_id
order by 1 desc,2 desc
```

Row	time_to_delivery	diff_estimated_delivery
1	0	28
2	0	26
3	0	20
4	0	17
5	0	13
6	0	12
7	0	12
8	0	12
9	0	10
10	0	10

2. Find time_to_delivery & diff_estimated_delivery. Formula for the same given below:
    - time_to_delivery = order_purchase_timestamp-order_delivered_customer_date
    - diff_estimated_delivery = order_estimated_delivery_date-order_delivered_customer_date

```
select date_diff(o.order_purchase_timestamp,
o.order_delivered_customer_date,day) as time_to_delivery,
date_diff(o.order_estimated_delivery_date,o.order_delivered_customer_date
,day) as diff_estimated_delivery from target.customers as cjoin target.
orders as o on c.customer_id=o.customer_id
```

Row	time_to_delivery	diff_estimated_delivery
1	-30	-12
2	-30	28
3	-35	16
4	-30	1
5	-32	0
6	-29	1
7	-43	-4
8	-40	-4

3. Group data by state, take mean of freight_value, time_to_delivery, diff_estimated_delivery

```
select AVG(oi.freight_value) avg_value,
avg(date_diff(o.order_purchase_timestamp,o.order_delivered_customer_date,day))
as time_to_delivery,
avg(date_diff(o.order_estimated_delivery_date,o.order_delivered_customer_date,
day)) as diff_estimated_delivery,c.customer_state
from target.customers as c join target.orders as o
on c.customer_id=o.customer_id join `target.order_items` as oi
on o.order_id=oi.order_id group by 4
```

Row	avg_value	time_to_delivery	diff_estimated_c	customer_state
1	28.1662843...	-17.5081967...	13.6393442...	MT
2	38.2570024...	-21.2037500...	9.10999999...	MA
3	35.8436711...	-23.9929742...	7.97658079...	AL
4	15.1472753...	-8.25960855...	10.2655943...	SP
5	20.6301668...	-11.5155221...	12.3971510...	MG
6	32.9178626...	-17.7920962...	12.5521191...	PE
7	20.9609239...	-14.6893821...	11.1444931...	RJ
8	21.0413549...	-12.5014861...	11.2747346...	DF
9	21.7358043...	-14.7082993...	13.2030001...	RS
10	36.6531688...	-20.9786666...	9.16533333...	SE

4. Sort the data to get the following:
5. Top 5 states with highest/lowest average freight value - sort in desc/asc limit 5

```
select AVG(oi.freight_value) avg_value,c.customer_state
from target.customers as c join target.orders as o
on c.customer_id=o.customer_id join `target.order_items` as oi
on o.order_id=oi.order_id group by 2
order by 1 desc limit 5
```

Row	avg_value	customer_state
1	42.9844230...	RR
2	42.7238039...	PB
3	41.0697122...	RO
4	40.0733695...	AC
5	39.1479704...	PI

6. Top 5 states with highest/lowest average time to delivery

```
select avg(date_diff(o.order_delivered_customer_date,o.order_purchase_timestamp
,day)) as time_to_delivery,c.customer_state
from target.customers as c join target.orders as o
on c.customer_id=o.customer_id
group by 2 order by 1 desc limit 5
```

Row	time_to_delivery	customer_state
1	28.9756097...	RR
2	26.7313432...	AP
3	25.9862068...	AM
4	24.0403022...	AL
5	23.3160676...	PA

7. Top 5 states where delivery is really fast/ not so fast compared to the estimated date

```
select avg(date_diff(o.order_purchase_timestamp,o.order_delivered_customer_date
,day)) as time_to_delivery,
avg(date_diff(o.order_estimated_delivery_date,o.order_delivered_customer_date,
day)) as diff_estimated_delivery,
c.customer_state from target.customers as c join target.orders as o
on c.customer_id=o.customer_id group by 3 order by 3 desc,1 desc limit 5
```

Row	time_to_delivery	diff_estimated_c	customer_state
1	-17.2262773...	11.2591240...	TO
2	-8.29806148...	10.1353253...	SP
3	-21.0298507...	9.17313432...	SE
4	-14.4795601...	10.6058641...	SC
5	-14.8192365...	12.9818488...	RS

6. Payment type analysis:

1. Month over Month count of orders for different payment types

```
select extract(month from o.order_purchase_timestamp) as month,
count(distinct(o.order_id))as Total_orders,
p.payment_type from target.orders as o
join target.payments as p
on o.order_id=p.order_id
group by 1,3
order by 1,2
```

Row	month	Total_orders	payment_type
1	1	118	debit_card
2	1	337	voucher
3	1	1715	UPI
4	1	6093	credit_card
5	2	82	debit_card
6	2	288	voucher
7	2	1723	UPI
8	2	6582	credit_card
9	3	109	debit_card
10	3	395	voucher

2. Count of orders based on the no. of payment installments

```
select count(distinct(o.order_id))as Total_orders,
p.payment_installments from target.orders as o
join target.payments as p on o.order_id=p.order_id
group by p.payment_installments order by 2
```

Row	Total_orders	payment_installments
1	2	0
2	49060	1
3	12389	2
4	10443	3
5	7088	4
6	5234	5
7	3916	6
8	1623	7
9	4253	8
10	644	9

1. Actionable Insights

    - Regarding the data the sales are rapidly increasing month – month.
    - Every year end of the year sales count is very high.
    - In 2017 from January to October the sales are slowing growing
    - So, try to put offers on products in the start of the month.
    - Based on data in a day morning and night period sales are less so keep offers on those periods.

2. Recommendations

- Based on data in a dawn period keep maintenance service because of less sales.
- In 2017 from January to October the sales are slowing growing