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

a)Data type of all columns in the "customers" table.

```
select column_name,data_type
from `target`.INFORMATION_SCHEMA.COLUMNS
where table_name= 'customers';
```

JOB INFORMATION | RESULTS | CHART | JSON

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.

```
select
min(order_purchase_timestamp) as mintime,
max(order_purchase_timestamp) as maxtime,
from `target.orders`
```

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

From the query result we can see that time range for the dataset is approximately 2 years which is from sep-16 to oct-18.

c) Count the Cities & States of customers who ordered during the given period.

```
SELECT
DISTINCT customers.customer_state,
customers.customer_city
FROM
`target.customers` customers
JOIN
`target.orders` orders
ON
customers.customer_id=orders.customer_id
ORDER BY
customers.customer_state;
```

Row	customer_state	customer_city
1	AC	xapuri
2	AC	brasileia
3	AC	porto acre
4	AC	rio branco
5	AC	manoel urbano
6	AC	epitaciolandia
7	AC	cruzeiro do sul
8	AC	senador guiomard
9	AL	belem
10	AL	igaci

2. In-depth Exploration

a)Is there a growing trend in the no. of orders placed over the past years?

```
SELECT
EXTRACT(year
FROM
order_purchase_timestamp) AS year,
EXTRACT(month
FROM
order_purchase_timestamp) AS month,
COUNT(order_id) AS order_count
FROM
`target.orders` orders
WHERE
orders.order_status="delivered"
GROUP BY
year,
month
ORDER BY
year,
month;
```

Row	year	month	order_count
1	2016	9	1
2	2016	10	265
3	2016	12	1
4	2017	1	750
5	2017	2	1653
6	2017	3	2546
7	2017	4	2303
8	2017	5	3546
9	2017	6	3135
10	2017	7	3872

Insights:

From the result we can see that there is a significant growth in count of orders from 2016 to 2017 but between 2017 to 2018 there is little dip in orders but again its increases toward the last of 2018.

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

```
SELECT
  EXTRACT(month
  FROM
    order_purchase_timestamp) AS month,
  COUNT(order_id) AS order_count
FROM
  `target.orders` orders
GROUP BY
  month
ORDER BY
  month;
```

Row	month	order_count
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

Insights:

Definitely there is monthly seasonality in terms of no. of order being placed with many to august month having highest volume of order followed by first four month and lowest order count is seen during September and October month.

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

```
SELECT
  time,
  COUNT(order_id) AS count_total
FROM (
  SELECT
    *,
    CASE
      WHEN EXTRACT(hour FROM order_purchase_timestamp) BETWEEN 0 AND 6 THEN
"Dawn"
      WHEN EXTRACT(hour
    FROM
      order_purchase_timestamp) BETWEEN 7
    AND 12 THEN "Mornings"
      WHEN EXTRACT(hour FROM order_purchase_timestamp) BETWEEN 13 AND 18
THEN "Afternoon"
      ELSE "Night"
  END
    time
  FROM
    `target.orders`)
GROUP BY
  time;
```

Row	time	count_total
1	Mornings	27733
2	Dawn	5242
3	Afternoon	38135
4	Night	28331

Insights:

From the results we can safely assume that Brazilian customers prefers to order during afternoon most followed by night and morning time and least preference for them is during dawn time.

Note:We as assumed time in the dataset as local Brazilian time.

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

a) Get the month on month no. of orders placed in each state.

```
SELECT
  EXTRACT(year
  FROM
    order_purchase_timestamp) year,
  EXTRACT(month
  FROM
    order_purchase_timestamp) month,
  ct.customer_state,
  COUNT(od.order_id) AS order_count
FROM
  `target.customers` ct
JOIN
  `target.orders` od
ON
  ct.customer_id=od.customer_id
GROUP BY
  ct.customer_state,
  year,
  month
ORDER BY
  ct.customer_state,
  year,
  month;
```

Row	year	month	customer_state	order_count
1	2017	1	AC	2
2	2017	2	AC	3
3	2017	3	AC	2
4	2017	4	AC	5
5	2017	5	AC	8
6	2017	6	AC	4
7	2017	7	AC	5
8	2017	8	AC	4
9	2017	9	AC	5
10	2017	10	AC	6

b) How are the customers distributed across all the states?

```
SELECT
  ct.customer_state,
  COUNT(customer_unique_id) AS customer_count
FROM
  `target.customers` ct
GROUP BY
  ct.customer_state
ORDER BY
  customer_count DESC
```

Row	customer_state	customer_count
1	SP	41746
2	RJ	12852
3	MG	11635
4	RS	5466
5	PR	5045
6	SC	3637
7	BA	3380
8	DF	2140
9	ES	2033
10	GO	2020

Insights:

By seeing the result we can say that highest no. of customers are from Sao paulo followed by Rio de Janeiro but having huge difference in numbers between them.so leaving the top 3 states the distribution of customers across all other states is not very high.so, we can infer that target is largely present in theses 3 states.

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

a) **Get the % increase in the cost of orders from year 2017 to 2018 (include months between Jan to Aug only).**

```
WITH
cte AS (
SELECT
EXTRACT(year
FROM
     od.order_purchase_timestamp) year,
SUM(py.payment_value) AS order_total
FROM
`target.orders` od
JOIN
`     target.payments` py
ON
od.order_id=py.order_id
WHERE
EXTRACT(month
FROM
     order_purchase_timestamp) BETWEEN 0
AND 8
AND od.order_status="delivered"
GROUP BY
year)
SELECT
ROUND(((c.order_total-ct.order_total)/ct.order_total)*100,2)
pct_increase
FROM
cte c
JOIN
cte ct
ON
c.year<>ct.year
ORDER BY
pct_increase DESC
LIMIT
1;
```

Row	pct_increase
1	143.33

There is 143% increase in cost of year from 2017 to 2018.

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

```
SELECT
customer_state,
ROUND(SUM(price),2) AS total_price,
ROUND(AVG(price),2) AS avg_price
FROM
`target.customers` ct
JOIN
target.orders od
ON
ct.customer_id=od.customer_id
JOIN
`target.order_items` odd
ON
od.order_id=odd.order_id
GROUP BY
customer_state
```

Row	customer_state	total_price	avg_price
1	RN	83034.98	156.97
2	CE	227254.71	153.76
3	RS	750304.02	120.34
4	SC	520553.34	124.65
5	SP	5202955.05	109.65
6	MG	1585308.03	120.75
7	BA	511349.99	134.6
8	RJ	1824092.67	125.12
9	GO	294591.95	126.27
10	MA	119648.22	145.2

Insights:

From the results we can infer that there is significant difference in total price and average price.
States like Sao paulo and Rio which have high customer count and higher total price have less average price but in states which have less total price have high average price signifying less customer purchasing less but high valued items.

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

```
SELECT
customer_state,
ROUND(SUM(freight_value),2) AS total_freight,
ROUND(AVG(freight_value),2) AS avg_freight
FROM
`target.customers` ct
JOIN
target.orders od
ON
ct.customer_id=od.customer_id
JOIN
`target.order_items` odd
ON
od.order_id=odd.order_id
GROUP BY
customer_state
```

Row	customer_state	total_freight	avg_freight
1	MT	29715.43	28.17
2	MA	31523.77	38.26
3	AL	15914.59	35.84
4	SP	718723.07	15.15
5	MG	270853.46	20.63
6	PE	59449.66	32.92
7	RJ	305589.31	20.96
8	DF	50625.5	21.04
9	RS	135522.74	21.74
10	SE	14111.47	36.65

Insights:

From results we can see that states like Sao and Rio have high freight value and low average freight value showing high quantity of logistic movement with lower cost spent on the shipping which indicates efficient logistics network in these large volume states.
On the other hand states like piaui have higher average freight value showing higher shipping and transportation cost as volume of orders is less.

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

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.

```
SELECT
order_id,
DATE_DIFF(order_delivered_customer_date,order_purchase_timestamp,day)
time_to_deliver,

 DATE_DIFF(order_delivered_customer_date,order_estimated_delivery_date
 ,day)    diff_estimated_delivery
FROM
`target.orders`
```

Row	order_id	time_to_deliver	diff_estimated_delive
1	1950d777989f6a877539f5379...	30	12
2	2c45c33d2f9cb8ff8b1c86cc28...	30	-28
3	65d1e226dfaeb8cdc42f66542...	35	-16
4	635c894d068ac37e6e03dc54e...	30	-1
5	3b97562c3aee8bdedcb5c2e45...	32	0
6	68f47f50f04c4cb6774570cfde...	29	-1
7	276e9ec344d3bf029ff83a161c...	43	4
8	54e1a3c2b97fb0809da548a59...	40	4
9	fd04fa4105ee8045f6a0139ca5...	37	1
10	302bb8109d097a9fc6e9cefc5...	33	5

Insights:

Time to deliver and estimated delivery difference can tell us about the efficiency of business and logistics and it can help business in taking right decisions.As there are some negative values in the estimated delivery difference volume it indicates early deliveries, which again can help business in deciding accurate timelines.

b) Find out the top 5 states with the highest & lowest average freight value.

```
(
SELECT
customer_state,
ROUND(AVG(freight_value),2) AS freight_value
FROM
`target.customers` ct
JOIN
target.orders od
ON
ct.customer_id=od.customer_id
JOIN
`target.order_items` odd
ON
od.order_id=odd.order_id
GROUP BY
customer_state
ORDER BY
freight_value DESC
LIMIT
5)
UNION ALL (
SELECT
customer_state,
ROUND(AVG(freight_value),2) AS freight_value
FROM
`target.customers` ct
JOIN
target.orders od
ON
ct.customer_id=od.customer_id
JOIN
`target.order_items` odd
ON
od.order_id=odd.order_id
GROUP BY
customer_state
ORDER BY
freight_value
LIMIT
5)
```

Row	customer_state	freight_value
1	RR	42.98
2	PB	42.72
3	RO	41.07
4	AC	40.07
5	PI	39.15
6	SP	15.15
7	PR	20.53
8	MG	20.63
9	RJ	20.96
10	DF	21.04

Insights:

States like Roraima have highest freight value showing very high logistics cost while on the other hand states like Sao paulo have lowest freight value showing very low logistics cost there.

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

```
(
SELECT
customer_state,

 ROUND(AVG(DATE_DIFF(order_delivered_customer_date,order_purchase_timesta
 mp,day    ) ),2)avg_time_to_deliver
FROM
`target.customers` ct
JOIN
target.orders od
ON
ct.customer_id=od.customer_id
GROUP BY
customer_state
ORDER BY
avg_time_to_deliver DESC
LIMIT
5)
UNION ALL (
SELECT
customer_state,

 ROUND(AVG(DATE_DIFF(order_delivered_customer_date,order_purchase_timesta
 mp,day    ) ),2)avg_time_to_deliver
FROM
`target.customers` ct
JOIN
target.orders od
ON
ct.customer_id=od.customer_id
GROUP BY
customer_state
ORDER BY
avg_time_to_deliver ASC
LIMIT
5)
```

Row	customer_state	avg_time_to_deliver
1	RR	28.98
2	AP	26.73
3	AM	25.99
4	AL	24.04
5	PA	23.32
6	SP	8.3
7	PR	11.53
8	MG	11.54
9	DF	12.51
10	SC	14.48

Insights:

From the result we can infer that state like Sao paulo take lowest average time to deliver order showcasing efficient logistics and transportation while state like Amapa which have very high avg delivery time have less efficient logistics and transportation and need to be worked on.

d) Find out the top 5 states where the order delivery is really fast as compared to the estimated date of delivery.

```
SELECT
customer_state,

 ROUND(AVG(DATE_DIFF(order_delivered_customer_date,order_estimated_deli
 very_date,day) ),2)diff_estimated_delivery
FROM
`target.customers` ct
JOIN
target.orders od
ON
ct.customer_id=od.customer_id
GROUP BY
customer_state
ORDER BY
diff_estimated_delivery ASC
LIMIT
5
```

Row	customer_state	diff_estimated_delive
1	AC	-19.76
2	RO	-19.13
3	AP	-18.73
4	AM	-18.61
5	RR	-16.41

Insights:

From the results acre(AC) have highest difference in actual and estimated delivery time indicating it to be the state having fastest delivery time followed by Rondonia(RO) and others.

6. Analysis based on the payments:

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

```
SELECT
EXTRACT(year
FROM
order_purchase_timestamp) year,
EXTRACT(month
FROM
order_purchase_timestamp) month,
payment_type,
COUNT(DISTINCT(pm.order_id)) total_count
FROM
`target.orders`od
JOIN
`target.payments` pm
ON
od.order_id=pm.order_id
GROUP BY
year,
month,
payment_type
ORDER BY
year,
month,
payment_type
```

Row	year	month	payment_type	total_count
1	2016	9	credit_card	3
2	2016	10	UPI	63
3	2016	10	credit_card	253
4	2016	10	debit_card	2
5	2016	10	voucher	11
6	2016	12	credit_card	1
7	2017	1	UPI	197
8	2017	1	credit_card	582
9	2017	1	debit_card	9
10	2017	1	voucher	33

Insights:

From the query we can see that there are various preferences for payment type for the customers but among all credit card emerges as clear winner showcasing its benefits, ease of use, installment preference and discounts.

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

```
SELECT
payment_installments,
COUNT(DISTINCT(order_id)) count_total
FROM
`target.payments`
WHERE
payment_installments>0
GROUP BY
payment_installments
ORDER BY
count_total DESC;
```

Row	payment_installment	count_total
1	1	49060
2	2	12389
3	3	10443
4	4	7088
5	10	5315
6	5	5234
7	8	4253
8	6	3916
9	7	1623
10	9	644

Recommendation:

- From the query we can see that most of Target customer is from Sao and Rio , so company should focus on expanding their business in other states and cities.
- Target should definitely focus on improving their logistics and transportation in low volume order state which will help in reducing average freight value and average product cost.
- Target should also focus on reducing delivery time in some states and cities having less customers for improving their experience which can help in gain of more customers.
- Company should provide offers, vouchers and discounts in states having less customer density to build a good customer base in areas where company have low visibility.
- Also company should focus on increasing sale of it’s Low price items which can help in increase order volumes in low customer density states.
- Target should offer memberships and loyalty programs to customers which will influence customers to buy more items at low cost.