SQL PROJECT: TARGET

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

1.Data type of all columns in the "customers" table.

Field name	Type
customer_id	STRING
customer_unique_id	STRING
customer_zip_code_prefix	INTEGER
customer_city	STRING
customer_state	STRING

2. Get the time range between which the orders were placed.

Ans=
```
select min(order_purchase_timestamp) as min_,
      max(order_purchase_timestamp) as max_
      from
`Target_database.orders`;
```

Query results

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS

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

observation:

As given in the context of problem statement time range from 2016 to 2018. To be precise we can say that time range is

given from 4 september 2016 to 17 oct 2018. only 4 months of 2016 is present in time range and approx 10 months of 2018.

3. Count the number of Cities and States in our dataset.

Ans=
```
select count(distinct(geolocation_city)) as city,
count(distinct(geolocation_state))as state
from `Target_database.geolocation`;
```

Query results

JOB INFORMATION | RESULTS | JSON

Row	city	state
1	8011	27

2. In-depth Exploration:

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

Ans=
```
select extract(year from
order_purchase_timestamp) as
year ,count(order_id) as count_order
from `Target_database.orders`
group by year
Order by year
```

Query results

Row	year	count_order
1	2016	329
2	2017	45101
3	2018	54011

Observation:

Yes there is a growing trend in the no. Of orders if we look at yearly wise there is steep increase .
As in 2016 only 4 months of time range for the no. Of orders so there is very minimum count of orders. as in initial period people don't know about the company and there is trust issue . time is needed to gain the trust of the customer so if we look at he year 2017 there is exponential growth in count of orders that,s same momentum in 2018 as only 8 months of time range in the year 2018.
It not fair to say that compare the count_of orders year wise as not all year has full months of data except 2017.

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

Ans=
```
select extract(month from
    order_purchase_timestamp) as month,
    count(order_id) as count_order
from `Target_database.orders`
group by month order by count_order desc
```

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

Observation:

yes there is monthly seasonality in the count of orders as in the 8th ,5th,7th, month there is (orders > 10000) so we say this is peak month by count of order and in the 9th and 10th month there is (order<5000) less count of orders in this month.

3.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 count(order_id) as order_count, case when
   extract(hour from
   order_purchase_timestamp)  between 0 and 6 then
   '0-6 hrs:Dawn'
 when extract(hour from
  order_purchase_timestamp)  between 7 and 12 then '7-
  12 hrs:Mornings'
 when extract(hour from
order_purchase_timestamp)  between 13 and 18 then
'13-18 hrs:Afternoon'
 when extract(hour from
 order_purchase_timestamp)  between 18 and 23 then
 '18-23 hrs:night'
 end as timing from `Target_database.orders`
 group by timing
 order by order_count;
```

Query result

Row	order_count	timing
1	5242	0-6 hrs:Dawn
2	27733	7-12 hrs:Mornings
3	28331	18-23 hrs:night
4	38135	13-18 hrs:Afternoon

Observation:

most of the order are from the afternoon so we can say that people order mostly their order At afternoon .

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

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

Ans=
```
select extract(month from
o.order_purchase_timestamp) as month,extract(year
from o.order_purchase_timestamp) as
year,count(o.order_id) as order_count,
c.customer_state as state
from `Target_database.customers` c left join
`Target_database.orders` o on
c.customer_id=o.customer_id
group by 1,2,4
order by 1,2
```

Query result:

Row	month	year	order_count	state
1	1	2017	299	SP
2	1	2017	108	MG
3	1	2017	54	RS
4	1	2017	97	RJ
5	1	2017	65	PR
6	1	2017	12	PA
7	1	2017	18	GO
8	1	2017	25	BA
9	1	2017	31	SC
10	1	2017	5	RN

3.2.How are the customers distributed across all the states?

Ans=
```
select count(customer_id) as
count_customer,customer_state from
`Target_database.customers`
group by customer_state
order by 1 desc
```

Query results

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

Observation:

Customers are distributed in very diverse order sau paulo and rio de jenerio has the highest no. Of customers in the state and roraima has the least no. Of customers in Brazil .if you see the difference in the maximum and minimum in term of no of customers that will be huge.

So company has to target the reason of state of why there are less customers in this state.

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

Ans=
```
select s.seller_state, sum(price) as total from
`Target_database.orders_items` o
inner join `Target_database.sellers` s
on o.seller_id=s.seller_id
group by s.seller_state
order by sum(price) desc
```

Query results

Row	seller_state	total
1	SP	8753396.210013...
2	PR	1261887.209999...
3	MG	1011564.740000...
4	RJ	843984.2200000...
5	SC	632426.0700000...
6	RS	378559.5400000...
7	BA	285561.5599999...
8	DF	97749.47999999...
9	PE	91493.84999999...

Observation:

Sau paulo has the highest value of order price in the state while acre has the lowest
Value of order price in brazil.

Avg value of order

Ans=
```
select s.seller_state, avg(price) as total from
`Target_database.orders_items` o
inner join `Target_database.sellers` s
on o.seller_id=s.seller_id
group by s.seller_state
order by avg(price) desc
```

Query results

Row	seller_state	avg_value
1	PB	449.8684210526...
2	BA	444.1081804043...
3	AM	392.3333333333...
4	RO	340.1571428571...
5	AC	267.0
6	CE	215.3259574468...
7	PI	210.1666666666...
8	PE	204.2273437499...
9	RN	178.4392857142...

Observation:

Pariba state has the highest avg value of order price while Maranhao has the lowest value of order price.

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

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

Ans=

```
with r1 as
        (select distinct * from (SELECT order_id,extract(month
    from order_purchase_timestamp ) as months,extract(year
    from  order_purchase_timestamp )as years FROM
    `Target_database.orders` ) as r
          where  months between 1 and 8
         order by years ,months ),
        r2 as
       (select years,months,sum(p.payment_value)as total from
r1 join `Target_database.payments` p on t1.order_id=p.order_id
         group by years,months
         order by years,months),
         r3 as
        (select *,lag(total)over(partition by years order by
years,months) as previous_total from r2
        order by years,months)
       select ,100((total-previous_total)/(previous_total))
from r3
```

2.Calculate the Total & Average value of order freight for each state.
Ans= total value of order freight of each state

```
 select s.seller_state, sum(freight_value)
as total_freight_value from
`Target_database.orders_items` o
inner join `Target_database.sellers` s
on o.seller_id=s.seller_id
group by s.seller_state
order by sum(freight_value) desc
```

Query results

Row	seller_state	total_freight_value
1	SP	1482487.669999...
2	MG	212595.0600000...
3	PR	197013.5200000...
4	SC	106547.0600000...
5	RJ	93829.89999999...
6	RS	57243.08999999...
7	BA	19700.68000000...
8	DF	18494.06000000...
9	GO	12565.49999999...

Observation:
“Sao paulo” has the highest total value of freight and “acre” has the least .

Avg value of freight in each state

Ans=
```
select s.seller_state, avg(freight_value) as
avg_freight_value from `Target_database.orders_items` o
inner join `Target_database.sellers` s
on o.seller_id=s.seller_id
group by s.seller_state
order by avg(freight_value) desc
```

Query result

Row	seller_state	avg_freight_value
1	RO	50.91285714285...
2	CE	46.38117021276...
3	PB	39.18815789473...
4	PI	36.94333333333...
5	AC	32.84
6	ES	32.71809139784...
7	MT	31.94296551724...
8	SE	31.849
9	BA	30.63869362363...

Observation:

"Rondonia" state has the highest avg freight value while "Sau paulo" has the lowest avg freight value.

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

1.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.

Ans=
```
select order_id,
       datetime_diff(order_delivered_customer_date,ord
    er _purchase_timestamp,day) as time_to_deliver,
    datetime_diff(order_estimated_delivery_date,order_
  delivered_customer_date,day) as
  diff_estimated_delivery
  from `Target_database.orders`
```

Query results

Row	order_id	time_to_deliver	diff_estimated_delivery
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

Observation:

Here (“-”) symbol in diff_estimated_delivery reflects the order is delivered late after whats the promised date.

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

States with highest average freight_value

Ans=
```
select s.seller_state as state, avg(freight_value)
as avg_freight_value from
`Target_database.orders_items` o
inner join `Target_database.sellers` s
on o.seller_id=s.seller_id
group by s.seller_state
order by avg(freight_value) desc limit 5
```

Query results

Row	state	avg_freight_value
1	RO	50.91285714285...
2	CE	46.38117021276...
3	PB	39.18815789473...
4	PI	36.94333333333...
5	AC	32.84

Observation:

Top states which has the highest value freight value are (“rondonia,ceara,paraiba,piani,acre).

States with lowest average freight_value

Ans=
```
select s.seller_state as state,
    avg(freight_value) as avg_freight_value from
    `Target_database.orders_items` o
   inner join `Target_database.sellers` s
  on o.seller_id=s.seller_id
  group by s.seller_state
  order by avg(freight_value) asc
  limit 5
```

Query results

Row	state	avg_freight_value
1	SP	18.45221266585...
2	PA	19.38874999999...
3	RJ	19.47486508924...
4	DF	20.57181312569...
5	PR	22.72096874639...

Observation:

Top 5 states with lowest average freight value are (“sao paulo, Para,Rio de jeneiro,Distrito federal,parana”).

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

Ans= top 5 states highest average delivery time

```
select
c.customer_state,avg(datetime_diff(order_delivered_cu
stomer_date, order_purchase_timestamp,day)) as
avg_delivery_time_in_days from
`Target_database.customers` c
inner join `Target_database.orders` o
on c.customer_id=o.customer_id
group by c.customer_state
order by avg_delivery_time_in_days desc
limit 5
```

Query results

Row	customer_state	avg_delivery_time_in_days
1	RR	28.975609756097562
2	AP	26.731343283582088
3	AM	25.986206896551735
4	AL	24.040302267002509
5	PA	23.316067653276953

Top 5 states with lowest delivery time

Ans=

```
select
c.customer_state,avg(datetime_diff(order_delive
red_customer_date,
order_purchase_timestamp,day)) as
avg_delivery_time_in_days from
`Target_database.customers` c
inner join `Target_database.orders` o
on c.customer_id=o.customer_id
group by c.customer_state
order by avg_delivery_time_in_days asc
limit 5
```

Query results

Row	customer_state	avg_delivery_time_in_days
1	SP	8.2980614890726656
2	PR	11.526711354864963
3	MG	11.543813298106565
4	DF	12.509134615384614
5	SC	14.479560191711288

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

Ans=

```
select
    c.customer_state,min(datetime_diff(order_delivered
    _customer_date, order_estimated_delivery_date,day))
    as delivery_time from `Target_database.customers`
    c
inner join `Target_database.orders` o
on c.customer_id=o.customer_id
group by c.customer_state
```

```
order by delivery_time  asc
limit 5
```

Query results

Row	customer_state	delivery_time
1	SP	-146
2	MA	-139
3	RS	-134
4	RJ	-108
5	MG	-77

Observation:

Sau paulo state has the highest delivery .here negative (-) sign indicates the product is delivered before the promised date.

6.Analysis based on the payments:

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

Ans=
```
select extract(month from o.order_purchase_timestamp) as month,extract(year from o.order_purchase_timestamp) as year,count(o.order_id) as order_count, p.payment_type from `Target_database.orders` o inner join
`Target_database.payments` p
on o.order_id=p.order_id
group by month,year,p.payment_type
order by 1,2;
```

Query results

Row	month	year	order_count	payment_type
1	1	2017	583	credit_card
2	1	2017	197	UPI
3	1	2017	61	voucher
4	1	2017	9	debit_card
5	1	2018	5520	credit_card
6	1	2018	1518	UPI
7	1	2018	416	voucher
8	1	2018	109	debit_card
9	2	2017	1356	credit_card
10	2	2017	398	UPI

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

Ans=
```
select count(order_id) as
    number_of_order ,payment_installments from
    `Target_database.payments`
    group by payment_installments;
```

Query result

Row	number_of_order	payment_installment
1	2	0
2	52546	1
3	12413	2
4	10461	3
5	7098	4
6	5239	5
7	3920	6
8	1626	7
9	4268	8
10	644	9

Observation:

In the number of order there is maximum order if the number of payment_installment is 1.

Actionable insights and recommendation:

Through the analyzing the dateset thoroughly penetration of e-commerce is not high as it is developing economy.some states has the very high count of orders and some has very low .and the difference in the top to bottom is huge .so it needs to be

filled with by knowing the customer needs as it changing from location to location as one specific location has the different needs from one another.
And the same with delivery time for some people it arrived before the estimated delivery date and for
Some long delay after the estimated delivery date.
Through the analytics of customer by doing so spending on the right area as it helps in bringing
the customer and rise the revenue of the company.
And one more thing that's the customer satisfaction with the company review score ,and any
Discomfort ,complain of the customer needs to be resolved in efficient way.