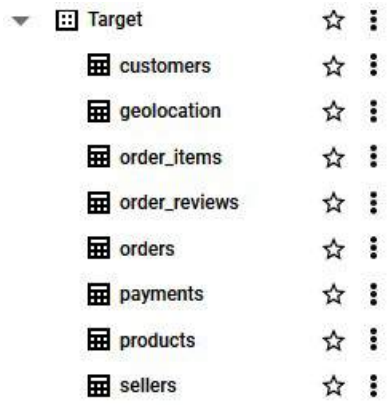
Exploratory analysis

26 July 2024 09:59

1. COLUMN DATA TYPE STUDY

Target dataset with all the tables



CUSTOMERS

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

GEOLOCATION

Field name	Type
geolocation_zip_code_prefix	INTEGER
geolocation_lat	FLOAT
geolocation_lng	FLOAT
geolocation_city	STRING
geolocation_state	STRING

ORDER_ITEMS

Field name	Type
order_id	STRING
order_item_id	INTEGER
product_id	STRING
seller_id	STRING
shipping_limit_date	TIMESTAMP
price	FLOAT
freight_value	FLOAT

ORDER REVIEWS

Field name	Type
review_id	STRING
order_id	STRING
review_score	INTEGER
review_comment_title	STRING
review_creation_date	TIMESTAMP
review_answer_timestamp	TIMESTAMP

ORDERS

Field name	Type
order_id	STRING
customer_id	STRING
order_status	STRING
order_purchase_timestamp	TIMESTAMP
order_approved_at	TIMESTAMP
order_delivered_carrier_date	TIMESTAMP
order_delivered_customer_date	TIMESTAMP
order_estimated_delivery_date	TIMESTAMP

PAYMENTS

Field name	Type
order_id	STRING
payment_sequential	INTEGER
payment_type	STRING
payment_installments	INTEGER
payment_value	FLOAT

PRODUCTS

Field name	Type
product_id	STRING
product_category	STRING

SELLERS

Field name	Type
seller_id	STRING
seller_zip_code_prefix	INTEGER
seller_city	STRING
seller_state	STRING

product_name_length	INTEGER
product_description_length	INTEGER
product_photos_qty	INTEGER
product_weight_g	INTEGER
product_length_cm	INTEGER
product_height_cm	INTEGER
product_width_cm	INTEGER

2. Query for timeframe of orders placed

```
select extract(year from order_purchase_timestamp) year,
extract(quarter from order_purchase_timestamp) quarter,
count(order_id) count_of_orders
from Target.orders
group by year,quarter
order by year, quarter
```

Table shows the timeframe of orders placed

Row	year	quarter	count_of_orders
1	2016	3	4
2	2016	4	325
3	2017	1	5262
4	2017	2	9349
5	2017	3	12642
6	2017	4	17848
7	2018	1	21208
8	2018	2	19979
9	2018	3	12820
10	2018	4	4

3. Query for number of customers from each state and city.

```
select c.customer_state,c.customer_city,
count(distinct c.customer_id) no_of_customers
from
Target.orders o join Target.customers c
on o.customer_id = c.customer_id
group by c.customer_state,c.customer_city
order by c.customer_state,c.customer_city
```

Sample output

Row	customer_state	customer_city	no_of_customers
1	AC	brasileia	1
2	AC	cruzeiro do sul	3
3	AC	epitaciolandia	1
4	AC	manoel urbano	1
5	AC	porto acre	1
6	AC	rio branco	70
7	AC	senador guiomard	2
8	AC	xapuri	2
9	AL	agua branca	1

Query for number of states and cities of customers who ordered during the given period.

```
select *,count(customer_state) over() total_no_of_states
from
(select c.customer_state, count(distinct c.customer_city) no_of_cities,
sum(count(distinct c.customer_city)) over() total_no_of_cities
from
Target.orders o join Target.customers c
on o.customer_id = c.customer_id
group by c.customer_state
)X
order by customer_state
```

Sample output

Row	customer_state	no_of_cities	total_no_of_cities	total_no_of_states
1	AC	8	4310	27
2	AL	68	4310	27
3	AM	5	4310	27
4	AP	6	4310	27
5	BA	353	4310	27
6	CE	161	4310	27
7	DF	6	4310	27
8	ES	95	4310	27
9	GO	178	4310	27

Results per page: 50 1 – 27 of 27

In-depth exploration

26 July 2024 10:05

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

Query

```
select extract (year from order_purchase_timestamp) year, count(order_id) no_of_orders
from
Target.orders
group by year
order by year
```

Sample output

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

Comments

Yes, there is large increase in number of orders from 2016 to 2017 and 2018. This may be because there is less data for the year 2016. There are no NULL values for order date and order_id.
There is 20% increase in number of orders during 2018 compared to 2017.

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

Query

```
select extract(year from order_purchase_timestamp) year, extract(month from
order_purchase_timestamp ) month, count(order_id) no_of_orders
from
Target.orders
group by year,month
order by year,month
```

Sample output

Row	year	month	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

Comments

In the year 2016 and 2018, there seems to be an anomaly as there are very less or no orders from the month of September to December. This may be because of no data. Another reason can be the stores were not operating on those months.

In the year there has been steady increase in number of orders.
So, there seems to be seasonality from september to december for year 2016 and year 2018. Year 2017 does not follow it.

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

Query

```
select
case when hour between 0 and 6
then 'dawn'
when hour between 7 and 12
then 'morning'
when hour between 13 and 18
then 'afternoon'
else
'night'
end period,
count(order_id) no_of_orders
from
(select extract(hour from order_purchase_timestamp) hour, order_id
from Target.orders
)
group by period
order by no_of_orders
```

Sample output

Row	period	no_of_orders
1	dawn	5242
2	morning	27733
3	night	28331
4	afternoon	38135

Comments

During afternoon the highest number of orders have been placed followed by night.

Evolution of E-commerce orders in the Brazil region

31 July 2024 23:40

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

Query

```
select *,
sum(no_of_orders) over(partition by customer_state) total_state_orders,
sum(no_of_orders) over(partition by month) total_monthly_orders
from
(
select c.customer_state, extract(month from o.order_purchase_timestamp) month,
count(o.order_id) no_of_orders
from
Target.orders o join Target.customers c
on o.customer_id = c.customer_id
group by c.customer_state,month) X
order by no_of_orders desc
```

Sample output

Row	customer_state	month	no_of_orders	total_state_orders	total_monthly_orders
1	SP	8	4982	41746	10843
2	SP	5	4632	41746	10573
3	SP	7	4381	41746	10318
4	SP	6	4104	41746	9412
5	SP	3	4047	41746	9893
6	SP	4	3967	41746	9343
7	SP	2	3357	41746	8508
8	SP	1	3351	41746	8069
9	SP	11	3012	41746	7544
10	SP	12	2357	41746	5674

Comments

The top 3 states in terms of number of orders are SP, RJ, MG. August, May, July are the top states for the same.

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

Query

```
select c.customer_state,
count(o.customer_id) no_of_customers
from
Target.orders o join Target.customers c
on o.customer_id = c.customer_id
group by c.customer_state
order by no_of_customers desc
```

Sample output

Row	customer_state	no_of_customers
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

Comments

The top three states in terms of number of customers are SP, RJ, MG. Consequently they are the states with highest number of orders.
It seems these states have customers with highest purchasing power. So, to increase the performance, Target has to target these customers and cater more to their needs.
To increase the sales in states with less customers, Target can reduce their pricing, give more discounts to cater to their low purchasing power customers.

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

01 August 2024 00:41

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

Query

```
select sum(p.payment_value) cost_of_order, extract(year from
o.order_purchase_timestamp) year
from
Target.payments p join Target.orders o
on
p.order_id = o.order_id
where (extract(year from o.order_purchase_timestamp) <> 2016) and
(extract(month from o.order_purchase_timestamp) between 1 and 8)
group by year
order by cost_of_order desc
```

Row	cost_of_order	year
1	8694733.839999...	2018
2	3669022.119999...	2017

Query for percentage increase in cost of orders

```
select
cost_of_order,
lead(cost_of_order) over(order by year desc) cost_of_order_prev,
round(
  ((cost_of_order-lead(cost_of_order) over(order by year desc))
        /lead(cost_of_order) over(order by year desc))*100,
    2
    ) percentage_increase,
year
from
(
select round(sum(p.payment_value),2) cost_of_order, extract(year from
o.order_purchase_timestamp) year
from
Target.payments p join Target.orders o
on
p.order_id = o.order_id
where (extract(year from o.order_purchase_timestamp) <> 2016) and
(extract(month from o.order_purchase_timestamp) between 1 and 8)
group by year
)X
order by year desc
```

Sample output

Row	cost_of_order	cost_of_order_prev	percentage_increase	year
1	8694733.84	3669022.12	136.98	2018
2	3669022.12	*null*	*null*	2017

Comments

There has been 137% increase in cost of orders. This is an exceptional growth. The company had done great business in 2018 compared to 2017. Customers are liking the products and the pricing

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

Query

```
select c.customer_state, round(sum(ot.price),2) sum_of_orderPrice, round(avg(ot.price),2)
avg_of_orderPrice
 from
 Target.order_items ot join Target.orders o
 on
 ot.order_id = o.order_id
 join Target.customers c
 on
 o.customer_id = c.customer_id
 group by c.customer_state
 order by sum_of_orderPrice desc,avg_of_orderPrice desc
```

Sample output

Row	customer_state	sum_of_orderPrice	avg_of_orderPrice
1	SP	5202955.05	109.65
2	RJ	1824092.67	125.12
3	MG	1585308.03	120.75
4	RS	750304.02	120.34
5	PR	683083.76	119.0
6	SC	520553.34	124.65
7	BA	511349.99	134.6
8	DF	302603.94	125.77
9	GO	294591.95	126.27
10	ES	275037.31	121.91

Comments

The total order value and avg order value analysis aligns with the previous analysis for total number of orders and customers for each state. The same states are in the top 3 places here as in previous analysis.
The average order value is higher for states with lower total order values. This maybe because there are less customers in those states or the average order value per customer can be higher.
For that we have to check the data for avg_orderPrice_perCust as shown below.

Query

```
select avg(avg_price) avg_orderPrice_perCust, max(count) max_no_order, customer_state
from
(select sum(ot.price) sum, avg(ot.price) avg_price, count(o.order_id) count,
c.customer_state,c.customer_id
from
Target.order_items ot join Target.orders o
on ot.order_id = o.order_id
join Target.customers c
on o.customer_id  = c.customer_id
group by c.customer_state,c.customer_id
order by avg_price desc, sum desc, count desc )X
group by customer_state
order by avg_orderPrice_perCust desc, max_No_order desc
```

Sample output

Avg_orderPrice_perCust represents average purchasing price of customers in each state
Max_no_order represents maximum number of orders placed by single customer in each

state.

Row	avg_orderPrice_perC	max_no_order	customer_state
1	201.7878665413...	6	PB
2	185.0354622871...	5	AL
3	184.3739917695...	6	AC
4	178.6287719298...	5	RO
5	172.3842113402...	6	PA
6	172.2054411764...	5	AP
7	167.3803285543...	6	TO
8	162.9763105117...	5	RN
9	160.4127822853...	5	PI
10	157.9097324792...	6	CE

max_no_order	customer_state
21	SP
20	GO
15	PR
12	SC
12	MG
11	BA
10	RJ
10	RS
6	PB
6	AC

Comments

PB,AL,AC has the highest purchasing power of customers. The customers do not buy in large quantities or come regularly to stores. But they order high value products.

SP,GO,PR has the most number of recurring customers coming to the stores. They have customers who buy often from their stores.

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

Query

```
select c.customer_state, sum(ot.freight_value) total_freight_value, avg(ot.freight_value)
avg_freight_value
from
Target.order_items ot join Target.orders o
on
ot.order_id = o.order_id
join Target.customers c
on
o.customer_id = c.customer_id
group by c.customer_state
order by total_freight_value desc, avg_freight_value desc
```

Sample output

Row	customer_state	total_freight_value	avg_freight_value
1	SP	718723.0699999...	15.14727539041...
2	RJ	305589.3100000...	20.96092393168...
3	MG	270853.4600000...	20.63016680630...
4	RS	135522.7400000...	21.73580433039...
5	PR	117851.6800000...	20.53165156794...
6	BA	100156.6700000...	26.36295892656...

6	BA	100156.0799999...	26.36395893656...
7	SC	89660.26000000...	21.47036877394...
8	PE	59449.65999999...	32.91786267995...
9	GO	53114.97999999...	22.76681525932...
10	DF	50625.49999999...	21.04135494596...

Comments

SP,RJ,MG are the top 3 states in terms of total freight value because of higher orders from these states.

Analysis based on sales, freight and delivery time.

02 August 2024 17:40

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.

Query

```
select order_id,
timestamp_diff(order_delivered_customer_date,order_purchase_timestamp,day)
delivery_time_days,
timestamp_diff(order_delivered_customer_date,order_estimated_delivery_date,day)
diff_deliveryTime_days
from
Target.orders
order by diff_deliveryTime_days desc, delivery_time_days
```

Sample output

Row	order_id	delivery_time_days	diff_deliveryTime_da
1	1b3190b2dfa9d789e1f14c05b...	208	188
2	ca07593549f1816d26a572e06...	209	181
3	47b40429ed8cce3aee9199792...	191	175
4	2fe324febf907e3ea3f2aa9650...	189	167
5	285ab9426d6982034523a855f...	194	166
6	440d0d17af552815d15a9e41a...	195	165
7	c27815f7e3dd0b926b5855262...	187	162
8	d24e8541128cea179a11a6517...	175	161
9	0f4519c5f1c541ddec9f21b3bd...	194	161
10	2d7561026d542c8dbd8f0daea...	188	159

Comments

Finding which states face late deliveries will give us insights about how to improve customer service which in turn will result in increase in sales.

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

Query

```
select customer_state,X.avg_freight_value
from
(select c.customer_state,
avg(ot.freight_value) avg_freight_value,
dense_rank() over(order by avg(ot.freight_value) desc) rank_
from
Target.order_items ot join Target.orders o
on
ot.order_id = o.order_id
join Target.customers c
on
o.customer_id = c.customer_id
group by c.customer_state
) X
```

```
where rank_ <= 5
order by X.avg_freight_value desc
```

Sample output

Top 5 states in terms of highest average freight value

Row	customer_state	avg_freight_value
1	RR	42.98442307692...
2	PB	42.72380398671...
3	RO	41.06971223021...
4	AC	40.07336956521...
5	PI	39.14797047970...

Top 5 states in terms of lowest average freight value

Row	customer_state	avg_freight_value
1	SP	15.14727539041...
2	PR	20.53165156794...
3	MG	20.63016680630...
4	RJ	20.96092393168...
5	DF	21.04135494596...

Comments

States with highest freight value maybe deliver products to far away locations as determined by the average delivery time shown below.

Row	customer_state	avg_freight_value	deliveryTime
1	SP	15.14727539041...	8.259608552419...
2	PR	20.53165156794...	11.48079306071...
3	MG	20.63016680630...	11.51552218007...
4	RJ	20.96092393168...	14.68938215750...
5	DF	21.04135494596...	12.50148619957...
6	SC	21.47036877394...	14.52098584675...
7	RS	21.73580433039...	14.70829936409...
8	ES	22.05877659574...	15.19280898876...
9	GO	22.76681525932...	14.94817742643...
10	MS	23.37488400488...	15.10727496917...

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

Query

```
select customer_state,X.avg_freight_value,X.deliveryTime
from
(select c.customer_state,
avg(ot.freight_value) avg_freight_value,
avg(timestamp_diff(o.order_delivered_customer_date,o.order_purchase_timestamp,day))
deliveryTime,
dense_rank() over(order by avg(timestamp_diff(o.order_delivered_customer_date,
o.order_purchase_timestamp,day)) desc) rank_
from
Target.order_items ot join Target.orders o
on
ot.order_id = o.order_id
```

```
join Target.customers c
on
o.customer_id = c.customer_id
group by c.customer_state
) X
where rank_ <= 5
order by X.deliveryTime desc
```

Sample output

Top 5 states in terms of highest average delivery time

Row	customer_state	avg_freight_value	deliveryTime
1	RR	42.98442307692...	27.82608695652...
2	AP	34.00609756097...	27.75308641975...
3	AM	33.20539393939...	25.96319018404...
4	AL	35.84367117117...	23.99297423887...
5	PA	35.83268518518...	23.30170777988...

Top 5 states in terms of lowest average delivery time

Row	customer_state	avg_freight_value	deliveryTime
1	SP	15.14727539041...	8.259608552419...
2	PR	20.53165156794...	11.48079306071...
3	MG	20.63016680630...	11.51552218007...
4	DF	21.04135494596...	12.50148619957...
5	SC	21.47036877394...	14.52098584675...

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

Query

```
select customer_state, avg_diff_deliveryTime_days
from
(select c.customer_state,
avg(timestamp_diff(o.order_estimated_delivery_date,o.order_delivered_customer_date,
day))avg_diff_deliveryTime_days,
dense_rank() over(order by avg(timestamp_diff(o.order_estimated_delivery_date,
o.order_delivered_customer_date,day)) desc) rank_
from
Target.orders o join Target.customers c
on
o.customer_id = c.customer_id
group by c.customer_state
  )X
where rank_ <= 5
order by avg_diff_deliveryTime_days desc
```

Sample output

Top 5 states in terms of fastest average delivery time compared to estimated delivery time

Row	customer_state	avg_diff_deliveryTim

1	AC	19.76250000000...
2	RO	19.13168724279...
3	AP	18.73134328358...
4	AM	18.60689655172...
5	RR	16.41463414634...

Top 5 states in terms of slowest average delivery time compared to estimated delivery time

Row	customer_state	avg_diff_deliveryTim
1	AL	7.947103274559...
2	MA	8.768479776847...
3	SE	9.173134328358...
4	ES	9.618546365914...
5	BA	9.934889434889...

Comments

States AC,RO,AP delivers their products really fast so that's why they charge higher freight value. All products are delivered well before estimated time.

Analysis based on the payments

02 August 2024 20:02

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

   Query

```
select p.payment_type,
extract(month from o.order_purchase_timestamp) month,
extract(year from o.order_purchase_timestamp) year,
count(p.payment_type) no_of_payments
from
Target.payments p join Target.orders o
on
p.order_id = o.order_id
group by
p.payment_type, year, month
order by
month,year
```

   Sample output

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

   Comments

   Customers prefer to use debit card as payment method as shown below. Voucher payments has not been popular among customers as there has been decrease in payments from them

   Data shows percentage increase/decrease of payment types from 2017 to 2018

Row	payment_type	per_increase_payme
1	debit_card	161.8483412322...
2	credit_card	21.40997454292...
3	UPI	7.414808582246...
4	voucher	-9.97687479352...

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

   Query

```
select count(*) no_of_orders_paidFull
from
(select order_id, sum(payment_value) total_payment
```

```
from
Target.payments
group by order_id
having
count(*) > 1
) p
join
Target.order_items ot
on
p.order_id = ot.order_id
where
p.total_payment = ot.price +ot.freight_value
```

Sample Output

Number of customers who opted for EMI payment and paid all the installments

Row	no_of_orders_paidFu
1	2010

Actionable Insights & Recommendations

02 August 2024 21:11

- Target's best customers come from states SP,RJ,MG in terms of number of customers, volume of orders and recurring visits to the stores. So they need to cater to their needs by offering more products and discounts.
- More number of stores can be opened in these states for sales growth.
- Most peolple tend pay via EMI payments. Target can offer interest free AMI options for low value products to attract customers.
- Target gets most of their business in evening and night periods. So it is optimum that stores operate at full capacity at these times.
- Some of the least products that target sell are childern clothing, PC games, insurance services can be explored in terms of competitors and market demand for increasing sales.