Business Case: Target SQL

Brief description:
Target sales data for Brazil. Data is from 2016` Sept to 2018` Oct.
It has info on sales and customer demographics

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.

Query:

```
SELECT table_name, column_name, data_type
FROM   `target_business_case`.INFORMATION_SCHEMA.COLUMNS
WHERE  table_name = 'customers'
```

Output:

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

Insights:

Customer table has 5 columns, with below mentioned data type for each column.

i. customer_id – String
ii. customer_unique_id – String
iii. customer_zip_code_prefix – 64 bit Integer
iv. customer_city – String
v. customer_state – String

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

Query:

```
SELECT
MIN(order_purchase_timestamp) as ORDERS_SATRT,
MAX(order_purchase_timestamp) as ORDERS_END
From `target_business_case.orders`
```

Output:

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

Insights:

Given data set has orders dated from 4th Sept' 2016 to 17th Oct' 2018.

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

Query:

```
Select
count(distinct customer_city) as city_count,
count(distinct customer_state) as state_count
From `target_business_case.customers`
```

Output:

Row	city_count	state_count
1	4119	27

Insights:

Given data set has orders from customers belonging to **4,119** cities of **27** different states.
This represents the geographical distribution of the customer orders, which can be used to understand the customer base in these regions better to make data driven decisions for marketing and logistic strategies.

2) In-depth Exploration:

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

Query-1:

```
Select
extract(year from order_purchase_timestamp) as year,
count(*) as no_of_orders
From `target_business_case.orders`
group by 1
order by 1
```

Output-1:

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

Insight-1:

Yes, there is clear indication of growth in orders each year.

But in above data set for 2016, 2018 data through out the year is not available.

So, lets write another query to get average orders per day in each year.

Query-2:

```
Select *,
      round(no_of_orders/days_per_year) as avg_order_per_day
from (Select
      extract(year from order_purchase_timestamp) as year,
      count(*) as no_of_orders,
      count(distinct extract(dayofyear from order_purchase_timestamp)) as
      days_per_year,
      From `target_business_case.orders`
      group by year
)
order by year
```

Output-2:

Row	year	no_of_orders	days_per_year	avg_order_per_day
1	2016	329	15	22.0
2	2017	45101	361	125.0
3	2018	54011	258	209.0

Insights:

Yes, there is clear increase in average orders per day over the years.

In 2016, average orders per day were **22** (Data we have is of 15 days)
In 2017, average orders per day were **125** (Data we have is of 361 days)
In 2018, average orders per day were **209** (Data we have is of 258 days)

Recommendation:

Data indicates clear increase in business activity and customer demand over time, which can be used for planning production and logistic strategies for next years.

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

Query:

```
Select *,
      lag(no_of_orders) over(order by year,month) as prev_month,
      no_of_orders - lag(no_of_orders) over(order by year,month) as
      difference_orders_per_month
From (Select
      extract(year from order_purchase_timestamp) as year,
      extract(month from order_purchase_timestamp) as month,
      count(*) as no_of_orders
      From `target_business_case.orders`
      group by year, month) t
order by year,month
```

Output:

Row	year	month	no_of_orders	prev_month	difference_orders_pr
1	2016	9	4	*null*	*null*
2	2016	10	324	4	320
3	2016	12	1	324	-323
4	2017	1	800	1	799
5	2017	2	1780	800	980
6	2017	3	2682	1780	902
7	2017	4	2404	2682	-278
8	2017	5	3700	2404	1296
9	2017	6	3245	3700	-455
10	2017	7	4026	3245	781
11	2017	8	4331	4026	305
12	2017	9	4285	4331	-46
13	2017	10	4631	4285	346
14	2017	11	7544	4631	2913
15	2017	12	5673	7544	-1871
16	2018	1	7269	5673	1596
17	2018	2	6728	7269	-541
18	2018	3	7211	6728	483
19	2018	4	6939	7211	-272
20	2018	5	6873	6939	-66
21	2018	6	6167	6873	-706
22	2018	7	6292	6167	125
23	2018	8	6512	6292	220
24	2018	9	16	6512	-6496
25	2018	10	4	16	-12

Insights:

Yes, there is certain seasonality in above data.

When we check differences in monthly orders data, we clearly see a reasonable dip (-ve change) in sales during months of

12 in 2016
4,6,12 in 2017
4,6 in 2018

So, Sales are peaking around March, May and November months in each Year.

Recommendation:

With above insight, Production and logistics can be planned accordingly during seasonal months for better efficiency.

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

Query:

```
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 'Morning'
    when extract(hour from order_purchase_timestamp) between 13 and 18 then 'Afternoon'
    when extract(hour from order_purchase_timestamp) between 19 and 23 then 'Night'
    else 'others'
  end as time_of_day,
  count(*) as order_count
From `target_business_case.orders`
group by time_of_day
order by order_count desc
```

Output:

Row	time_of_day	order_count
1	Afternoon	38135
2	Night	28331
3	Morning	27733
4	Dawn	5242

Insights:

According to above query result, most of the orders by Brazilian customers are placed in **Afternoon**.

Recommendation:

With above insight, Store/Delivery man power management is recommended to plan for higher number of orders in the afternoon. Thus, optimizing the workforce cost and efficiency planning with data driven approach.

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

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

Query:

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

Output:

Row	customer_state	year	month	order_count
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
11	AC	2017	11	5
12	AC	2017	12	5
13	AC	2018	1	6
14	AC	2018	2	3
15	AC	2018	3	2
16	AC	2018	4	4
17	AC	2018	5	2
18	AC	2018	6	3
19	AC	2018	7	4

Insights:

Insights can be drawn for seasonality for each state from above output.
With this increased visibility of data for each state, Insights can be drawn to find geographical distribution of the sales and business.

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

Query:

```
select customer_state,
  count(customer_id) as customer_count
from `target_business_case.customers`
group by customer_state
order by customer_count desc
```

Output:

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
11	PE	1652
12	CE	1336
13	PA	975
14	MT	907
15	MA	747
16	MS	715
17	PB	536
18	PI	495
19	RN	485

Insights:

Highest number of customers are in state of SP (São Paulo) followed by RJ (Rio de Janeiro).

Recommendation:

Customer's geographical insights from above data can be used to plan marketing cost/plan strategies to efficiently increase the customer base.

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). You can use the "payment_value" column in the payments table to get the cost of orders.

Query:

```
Select round(100*(total_2018 - total_2017)/total_2017,2) as percentage_change
from ( select
        sum ( case
              when extract(year from order_purchase_timestamp) = 2017 and
                    extract(month from order_purchase_timestamp) between 1 and 8
              then p.payment_value
              else 0 end
            ) total_2017,
        sum ( case
                when extract(year from order_purchase_timestamp) = 2018 and
                      extract(month from order_purchase_timestamp) between 1 and 8
                then p.payment_value
                else 0 end) total_2018
from `target_business_case.orders` o left join `target_business_case.payments` p
on o.order_id = p.order_id
)temp
```

Output:

Row	percentage_change
1	136.98

Insights:

There is 136.98% increase in cos of order for Jan – Aug period of 2017 and 2018. This insight crucial for the company to assess the growth in business revenues and expenses.

Recommendation:

Analysing these factors can help Target plan on budgeting, financial, pricing strategies.

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

Query:

```
select
c.customer_state,
round(sum(oi.price),2) as total_price,
round(avg(oi.price),2) as avg_price
from `target_business_case.order_items` oi left join `target_business_case.orders`
o on oi.order_id = o.order_id
left join
`target_business_case.customers` c on o.customer_id = c.customer_id
group by c.customer_state
order by c.customer_state
```

Output:

Row	customer_state	total_price	avg_price
1	AC	15982.95	173.73
2	AL	80314.81	180.89
3	AM	22356.84	135.5
4	AP	13474.3	164.32
5	BA	511349.99	134.6
6	CE	227254.71	153.76
7	DF	302603.94	125.77
8	ES	275037.31	121.91
9	GO	294591.95	126.27
10	MA	119648.22	145.2
11	MG	1585308.03	120.75
12	MS	116812.64	142.63
13	MT	156453.53	148.3
14	PA	178947.81	165.69
15	PB	115268.08	191.48
16	PE	262788.03	145.51

Insights:

We have multiple insights from above data.
SP & RJ states have higher total order values, which reflects on the volumes of business concentrated in these states.

PB & AL states have highest average price per order, this signifies that people in these states are potentially high valued customers but still no. of customers/ orders is not up to with high volume business states.

Recommendation:

With above insights, Strategies/Plans are to be made to improve according the specific geographical distribution of customers like,

Increase the customer base with high average order value, as we have high purchase power customers here.

Plan the logistics/inventory according to volume of sales in each state.

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

Query:

```
select c.customer_state,
        round(sum(oi.freight_value),2) as total_price,
        round(avg(oi.freight_value),2) as avg_price
from `target_business_case.order_items` oi
left join `target_business_case.orders` o
on oi.order_id = o.order_id
left join `target_business_case.customers` c
on o.customer_id = c.customer_id
group by c.customer_state
order by c.customer_state
```

Output:

Row	customer_state	total_price	avg_price
1	AC	3686.75	40.07
2	AL	15914.59	35.84
3	AM	5478.89	33.21
4	AP	2788.5	34.01
5	BA	100156.68	26.36
6	CE	48351.59	32.71
7	DF	50625.5	21.04
8	ES	49764.6	22.06
9	GO	53114.98	22.77
10	MA	31523.77	38.26
11	MG	270853.46	20.63
12	MS	19144.03	23.37
13	MT	29715.43	28.17
14	PA	38699.3	35.83
15	PB	25719.73	42.72
16	PE	59449.66	32.92
17	PI	21218.2	39.15

Insights:

SP states has the highest total freight cost, indicating that most of the shipments are sent to or originate from this state.

Recommendation:

With such insights on freight costs and budget, plans should be made by company to optimize its shipping routes, new ware houses, negotiate better logistic vendor deals with carriers and warehouse facilitators thus trying to reduce transportation expenses.

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.

Query:

```
select order_id,
TIMESTAMP_DIFF(order_delivered_customer_date,order_purchase_timestamp,day) as
deliver_time,
TIMESTAMP_DIFF(order_estimated_delivery_date,order_delivered_customer_date,day)
as diff_in_estimated_delivery
from `target_business_case.orders`
order by order_id
```

Output:

Row	order_id	deliver_time	diff_in_estimated_de
1	00010242fe8c5a6d1ba2dd792...	7	8
2	00018f77f2f0320c557190d7a...	16	2
3	000229ec398224ef6ca0657da...	7	13
4	00024acbcdf0a6daa1e931b03...	6	5
5	00042b26cf59d7ce69dfabb4e...	25	15
6	00048cc3ae777c65dbb7d2a06...	6	14
7	00054e8431b9d7675808bcb8...	8	16
8	000576fe39319847cbb9d288c...	5	15
9	0005a1a1728c9d785b8e2b08...	9	0
10	0005f50442cb953dcd1d21e1f...	2	18
11	00061f2a7bc09da83e415a52d...	4	10
12	00063b381e2406b52ad42947...	10	0
13	0006ec9db01a64e59a68b2c34...	6	21
14	0008288aa423d2a3f00fcb17c...	12	7
15	0009792311464db532ff765bf...	7	5

Insights:

Difference is estimated delivery date and actual delivery data will provide insights to, inspect the reasons for delays in the delivery.
Analysing the pattern of delays over logistic warehouses/transports to fix them and avoid these cases.

Recommendation:

Analys the delays in delivery, if a pattern is found toward a logistic node/ transport vendor. This provide data point to change/renegotiate logistics strategies.

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

Highest average freight value:

Query:

```
select c.customer_state,
round(avg(oi.freight_value),2) as avg_price
from `target_business_case.order_items` oi
left join `target_business_case.orders` o
on oi.order_id = o.order_id
left join `target_business_case.customers` c
on o.customer_id = c.customer_id
group by c.customer_state
order by avg_price desc
limit 5
```

Output:

Row	customer_state	avg_price
1	RR	42.98
2	PB	42.72
3	RO	41.07
4	AC	40.07
5	PI	39.15

Lowest average freight value:

Query:

```
select c.customer_state,
round(avg(oi.freight_value),2) as avg_price
from `target_business_case.order_items` oi
left join `target_business_case.orders` o
on oi.order_id = o.order_id
left join `target_business_case.customers` c
on o.customer_id = c.customer_id
group by c.customer_state
order by avg_price
limit 5
```

Output:

Row	customer_state	avg_price
1	SP	15.15
2	PR	20.53
3	MG	20.63
4	RJ	20.96
5	DF	21.04

Insights:

The states with the highest average freight values are RR, PB, RO, AC, PI. Which are mostly either in North East or North Western part of Brazil.
The states with lowest average freight values are SP, PR, MG, RJ, DF which are mostly in south eastern part of Brazil.

Recommendation:

Above Insight clearly says that, Target has to improve the strategies for shipping in Northern part of the Brazil country. With decisions like setting up new logistic hub or partnering with new vendors in northern part of Brazil.

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

Highest average delivery time:

Query:

```
select c.customer_state,
round(avg(TIMESTAMP_DIFF(order_delivered_customer_date,order_purchase_tim
estamp,day))) as avg_delivery_time
from `target_business_case.order_items` oi
left join `target_business_case.orders` o
on oi.order_id = o.order_id
left join `target_business_case.customers` c
on o.customer_id = c.customer_id
group by c.customer_state
order by avg_delivery_time desc
limit 5
```

Output:

Row	customer_state	avg_delivery_time
1	RR	28.0
2	AP	28.0
3	AM	26.0
4	AL	24.0
5	PA	23.0

Lowest average delivery time:

Query:

```
select c.customer_state,
round(avg(TIMESTAMP_DIFF(order_delivered_customer_date,order_purchase_tim
estamp,day))) as avg_delivery_time
from `target_business_case.order_items` oi
left join `target_business_case.orders` o
on oi.order_id = o.order_id
left join `target_business_case.customers` c
on o.customer_id = c.customer_id
group by c.customer_state
order by avg_delivery_time
limit 5
```

Output:

Row	customer_state	avg_delivery_time
1	SP	8.0
2	PR	11.0
3	MG	12.0
4	DF	13.0
5	SC	15.0

Insights:

The states with the highest average freight values are RR, AP, AM, AL, PA. Which are mostly either in North East or North Western part of Brazil.
The states with lowest average freight values are SP, PR, MG, DF, SC which are mostly in south eastern part of Brazil.

Recommendation:

Above Insight clearly says that, Target has to improve the strategies for shipping in Northern part of the Brazil country. With decisions like setting up new logistic hub or partnering with new vendors in northern part of Brazil to achieve faster shipping. Which improves customer satisfaction.

4. 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.

Query:

```
select c.customer_state,
round(avg(TIMESTAMP_DIFF(order_estimated_delivery_date,
          order_delivered_customer_date,day))) as diff_estimated_delivery
from `target_business_case.order_items` oi
left join `target_business_case.orders` o
on oi.order_id = o.order_id
left join `target_business_case.customers` c
on o.customer_id = c.customer_id
group by c.customer_state
order by diff_estimated_delivery
limit 5
```

Output:

Row	customer_state	diff_estimated_deliv
1	AL	8.0
2	SE	9.0
3	MA	9.0
4	BA	10.0
5	ES	10.0

Insights:

These are high-performing states, to optimise delivery processes the company should elevate its overall delivery performance, increase customer satisfaction, and solidify its position as a reliable and efficient provider in the market.

Recommendation:

Practices and process followed at these high performing states to be captured and compared with lo performing states. Thus finding ways to improve the low performing states management efficiency.

6) Analysis based on the payments:

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

Query:

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

Output:

Row	year	month	payment_type	payment_type_order
1	2016	9	credit_card	3
2	2016	10	UPI	63
3	2016	10	credit_card	254
4	2016	10	debit_card	2
5	2016	10	voucher	23
6	2016	12	credit_card	1
7	2017	1	UPI	197
8	2017	1	credit_card	583
9	2017	1	debit_card	9
10	2017	1	voucher	61
11	2017	2	UPI	398
12	2017	2	credit_card	1356
13	2017	2	debit_card	13
14	2017	2	voucher	119
15	2017	3	UPI	590

Insights:

By closely monitoring the month-on-month trends of payment methods, the company can adapt its strategies to align with customer preferences, boost customer satisfaction, and maximise sales opportunities.

Recommendation:

Company should invest in payment technologies, renegotiate payment gateway/ payment vendor fee. Adapt new payment technologies which is also useful, easy to use, preferrable for customer.

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

Query:

```
select count(distinct order_id) as payment_order_count
from `target_business_case.payments`
where
payment_installments > 0
```

Output:

Row	payment_order_coun
1	99438

Insights:

This data provides valuable insights into customer behaviour and their willingness to use installment-based payment options.

Recommendation:

With the number of customers opting for installment payments, the company should ensure that its payment infrastructure is robust and capable of handling increased transaction volumes.