Business Case: Target SQL

Submitted by: Anjali Jaitly

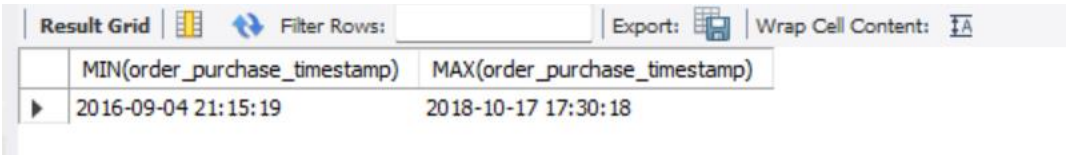
1. **Import the dataset and do usual exploratory analysis steps like checking the structure & characteristics of the dataset**
    1. Data type of columns in a table

Table	Datatype
products	TEXT, INT
order_items	TEXT, INT, DOUBLE
orders	TEXT
order_reviews	TEXT
customers	TEXT
sellers	TEXT
geolocation	TEXT, DOUBLE
payments	TEXT, INT, DOUBLE

    2. Time period for which the data is given

    QUERRY:

```
SELECT
        MIN(order_purchase_timestamp),
         MAX(order_purchase_timestamp)
 FROM orders;
```



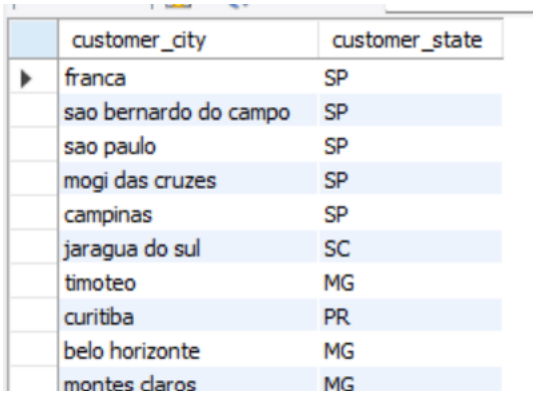
    INSIGHT : *The data is available from 4th September 2016 to 17th October 2018.*

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

    QUERRY:

```
SELECT
        DISTINCT
        customer_city,
         customer_state
```

```
FROM
    customers;
```

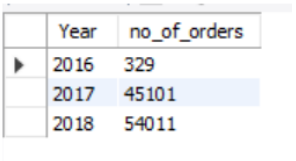


customer_city	customer_state
franca	SP
sao bernardo do campo	SP
sao paulo	SP
mogi das cruzes	SP
campinas	SP
jaragua do sul	SC
timoteo	MG
curitiba	PR
belo horizonte	MG
montes claros	MG

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?

QUERRY:

```
SELECT
    YEAR(order_purchase_timestamp) AS Year,
    count(*) AS no_of_orders
FROM
    orders
GROUP BY
    Year
ORDER BY
    Year;
```



Year	no_of_orders
2016	329
2017	45101
2018	54011

INSIGHT : *The number of orders placed in 2016 is only 329 as the data is available for only last for months from September to December. The number of orders placed has significantly increased from 45101 in 2017 to 54011 in 2018. This indicates a growing trend in e-commerce in Brazil.*

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

QUERRY:

```
SELECT
        time_of_day,
        count(*) AS num_of_orders
FROM
        (SELECT
                CASE
                        WHEN HOUR(order_purchase_timestamp) BETWEEN 0
AND 6
                 THEN "DAWN"
                WHEN HOUR(order_purchase_timestamp) BETWEEN 7 AND 12
                THEN "MORNING"
                 WHEN HOUR(order_purchase_timestamp) BETWEEN 12 AND 17
                THEN "AFTERNOON"
                WHEN HOUR(order_purchase_timestamp) BETWEEN 18 AND 24
                THEN "NIGHT"
                        END AS time_of_day
        FROM
                orders) AS t
GROUP BY
        time_of_day
ORDER BY
    num_of_orders DESC;
```

time_of_day	num_of_orders
NIGHT	34100
AFTERNOON	32366
MORNING	27733
DAWN	5242

INSIGHT : *Most of the purchases by Brazilian customers were made at Night between 5:00 pm and Midnight, followed by Afternoon and Morning. Least purchases were made at Dawn.*

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

1. Get month on month orders by states

QUERRY:

```
SELECT
```

```
        c.customer_state,
         MONTH(o.order_purchase_timestamp) AS Month,
        count(*) AS num_of_purchase
FROM
        orders AS o
JOIN
        customers AS c
ON
        o.customer_id = c.customer_id
GROUP BY
        customer_state, Month
ORDER BY
        customer_state, Month;
```

customer_state	Month	num_of_purchase
AC	1	2
AC	2	2
AC	3	1
AC	4	4
AC	5	4
AC	6	3
AC	7	2
AC	8	4
AC	9	2

INSIGHT : *People from SP state have made maximum purchases all the year round. The maximum orders have been placed from May to August in most of the states.*

2. Distribution of customers across the states in Brazil

   QUERRY:

```
SELECT
        customer_state,
        count(DISTINCT customer_id) AS num_of_customers
FROM
        customers
GROUP BY
         customer_state
ORDER BY
        num_of_customers DESC;
```

customer_state	num_of_customers
SP	14148
RJ	4459
MG	3970
RS	1850
PR	1715
SC	1201
BA	1133
DF	721
ES	692

INSIGHT : *The maximum 14148 of customers are from the state of SP, followed by RJ and MG. Minimum 20 customers are from the state of RR.*

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

QUERRY:

```
WITH table1 AS (
SELECT
        YEAR(o.order_purchase_timestamp) AS Year,
        SUM(p.payment_value) AS cost_per_year
FROM
        orders AS o
JOIN
        payments AS p
ON
        o.order_id = p.order_id
WHERE
        month(o.order_purchase_timestamp) BETWEEN 1 AND 8
GROUP BY
        year
)
SELECT
        ROUND((year_2018 - year_2017) / year_2017 * 100, 2) AS
        percent_increase
FROM
        (SELECT
                SUM(CASE WHEN Year = "2018" THEN
                cost_per_year END) AS year_2018,
```

```
        SUM(CASE WHEN Year = "2017" THEN
        cost_per_year END) AS year_2017
    FROM
        table1) AS t
```

percent_increase
136.98

INSIGHT : *There was 136.98% increase in cost of orders from 2017 to 2018.*

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

QUERRY:

```
    SELECT
        c.customer_state,
        ROUND(AVG(ot.price + ot.freight_value), 2) AS
mean_price,
        ROUND(SUM(ot.price + ot.freight_value), 2) AS
sum_price
    FROM
        customers AS c
    JOIN
        orders AS o
    ON
        c.customer_id = o.customer_id
    JOIN
        order_items AS ot
    ON
        o.order_id = ot.order_id
    GROUP BY
        c.customer_state
    ORDER BY
        mean_price DESC, sum_price DESC;
```

customer_state	mean_price	sum_price
RO	246.6	20714.74
AL	242.79	33505.28
PB	218.25	37539.22
PI	216.35	32236.62
AM	204.02	7752.61
TO	200.73	21076.52
RR	192.14	3458.46
SE	190.91	24626.9
MT	187.99	63353.32

INSIGHT : *The state of RO had the highest mean price followed by state by AL.*

5. Analysis on sales, freight and delivery time

1. Calculate days between purchasing, delivering and estimated delivery
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
3. Group data by state, take mean of freight_value, time_to_delivery, diff_estimated_delivery
4. Sort the data to get the following:
5. Top 5 states with highest/lowest average freight value - sort in desc/asc limit 5
6. Top 5 states with highest/lowest average time to delivery
7. Top 5 states where delivery is really fast/ not so fast compared to estimated date

TOP 5 states with highest average freight value

QUERRY:

```
SELECT
        c.customer_state,
        ROUND(AVG(ot.freight_value),2) AS mean_freight_value
FROM
        customers AS c
OIN
        orders AS o
ON
        c.customer_id = o.customer_id
JOIN
        order_items AS ot
ON
        o.order_id = ot.order_id
GROUP BY
        c.customer_state
ORDER BY
        mean_freight_value DESC
LIMIT
        5;
```

customer_state	mean_freight_val
RO	43.01
RR	41.02
AL	40.95
PB	40.92
PI	40.64

INSIGHT : *The top 5 states with highest mean freight value are RO, RR, AL, PB and* PI.

TOP 5 states with lowest average freight value

QUERRY:

```
SELECT
        c.customer_state,
        ROUND(AVG(ot.freight_value),2) AS mean_freight_value
FROM
        customers AS c
JOIN
        orders AS o
ON
        c.customer_id = o.customer_id
JOIN
        order_items AS ot
ON
        o.order_id = ot.order_id
GROUP  BY
        c.customer_state
ORDER BY
        mean_freight_value
LIMIT
        5;
```

customer_state	mean_freight_value
SP	15.13
RJ	20.6
MG	20.66
PR	21.09
SC	21.32

INSIGHT : *Top 5 states with lowest mean freight value are SP, RJ, MG, PR, SC.*

Top 5 states with highest with time to delivery

QUERRY:

```
SELECT
        c.customer_state,
        AVG(DATEDIFF(order_delivered_customer_date,
        order_purchase_timestamp)) AS mean_time_to_delivery
FROM
        customers AS c
```

```
JOIN
        orders AS o
ON
        c.customer_id = o.customer_id
GROUP BY
        c.customer_state
ORDER BY
        mean_time_to_delivery DESC
LIMIT
        5
```

customer_state	mean_time_to_deliver
AM	26.1220
RR	25.0000
AL	24.7200
PA	23.5802
AP	22.6250

INSIGHT : *Top 5 states with highest mean time to delivery are AM, RR, AL, PA and AP*

Top 5 states with lowest with time to delivery

QUERRY:

```
SELECT
        c.customer_state,
        AVG(DATEDIFF(order_delivered_customer_date,
        order_purchase_timestamp)) AS mean_time_to_delivery
FROM
        customers AS c
JOIN
        orders AS o
ON
        c.customer_id = o.customer_id
GROUP BY
        c.customer_state
ORDER BY
        mean_time_to_delivery
LIMIT
        5
```

customer_state	mean_time_to_delivery
SP	8.6905
MG	11.8729
PR	12.1118
DF	12.9428
RJ	15.0252

INSIGHT : *Top 5 states with lowest mean time to delivery are SP, MG, PR, DF and RJ.*

6. Payment type analysis:

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

QUERRY:

```
SELECT
        YEAR(o.order_purchase_timestamp) AS Year,
        MONTH(o.order_purchase_timestamp) AS Month,
        p.payment_type,
        COUNT(*) AS num_of_orders
FROM
        payments AS p
JOIN
        orders AS o
ON p.order_id = o.order_id
GROUP BY
        year, month, p.payment_type
ORDER BY
        year, month, p.payment_type
```

Year	Month	payment_type	num_of_orders
2016	9	credit_card	3
2016	10	credit_card	254
2016	10	debit_card	2
2016	10	UPI	63
2016	10	voucher	23
2016	12	credit_card	1
2017	1	credit_card	583
2017	1	debit_card	9
2017	1	UPI	197

INSIGHT : *The results indicate that month after month the most used payment method was credit card, followed by UPI.*

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

QUERRY:

```
SELECT
        payment_installments,
        count(*) AS num_of_orders
FROM
        payments
```

```
GROUP BY
        payment_installments
ORDER BY
        num_of_orders DESC
```

payment_installments	num_of_orders
1	52546
2	12413
3	10461
4	7098
10	5328
5	5239
8	4268
6	3920
7	1626

INSIGHT : *Maximum number of payments were done in single instalment. A significant number were done in 2 and 3 instalments as well. There was one payment which was even paid in 23 instalments and one in 22 instalments.*

RECOMMENDATIONS:

1. *Maximum number of customers for Target in Brazil exist in SP followed by MG and PR. The number of orders has also been placed in the same order, maximum from SP followed by MG and PR. This has a relation with the mean time to delivery which is least for SP next for MG and PR. Keeping this in mind if the delivery time to other states can be reduced, there might be an increase in number of customers from other states as well.*
2. *The maximum number of orders have been made at the night time, therefore it is recommended that Target makes more offers during this time to increase the sales.*
3. *May to August is the time of the year when most orders are placed in Brazil. The maximum offers and sales can be placed during this period.*