Business Case: Target SQL

Tool used : BigQuery
Dataset : Target
Analyzed by Rebekha Thangamani

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**

Below is the example query to get the data types of the column from a specific table in the dataset. 'Target' is my Dataset name and using customers table as an example.

Query:

```
select column_name, data_type from `Target.INFORMATION_SCHEMA.COLUMNS`

where table_name = 'customers';
```

Sample result:

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

INFORMATION_SCHEMA.COLUMNS - Helps to get all the information about the column.

(SELECT * FROM Target.INFORMATION_SCHEMA.TABLES;) this query helps to get all the information from the dataset like no.of tables, name of the tables, etc.

2.Time period for which the data is given:

As per the given context this business case has information of 100k orders from 2016 to 2018 made at Target in Brazil. The first order started on ***2016-09-04*** and the last orders was placed on ***2018-10-17***

Query:

```
select min(order_purchase_timestamp) as First_order_date,
max(order_purchase_timestamp) as last_order_date from Target.orders;
```

Result:

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

3.Cities and States covered in the dataset

(i) Below is the query to extract the states and cities of the customers based on their orders.

Query:

```
select DISTINCT customer_city, customer_state from Target.customers;
```

Result: (showing only 10 rows)

Row	customer_city	customer_state
1	acu	RN
2	ico	CE
3	ipe	RS
4	ipu	CE
5	ita	SC
6	itu	SP
7	jau	SP
8	luz	MG
9	poa	SP
10	uba	MG

(ii) If you want to know all the cities and states covered in the Dataset. We can use a similar query in the geolocation table.

```
select DISTINCT geolocation_city, geolocation_state from Target.geolocation;
```

2.In-depth Exploration:

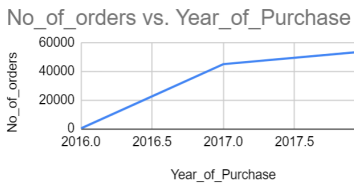
1. **Is there a growing trend on e-commerce in Brazil? How can we describe a complete scenario? Can we see some seasonality with peaks at specific months?**

Query: observation 1

```
SELECT * FROM(
SELECT count(order_id) No_of_orders, Extract(YEAR FROM order_purchase_timestamp) as Year_of_Purchase
from Target.orders
GROUP BY Extract(YEAR FROM order_purchase_timestamp))t
ORDER BY Year_of_Purchase
```

Result: **We can see the number of orders placed by the customers increased rapidly over years.**

No_of_orders	Year_of_Purchase
329	2016
45101	2017
54011	2018

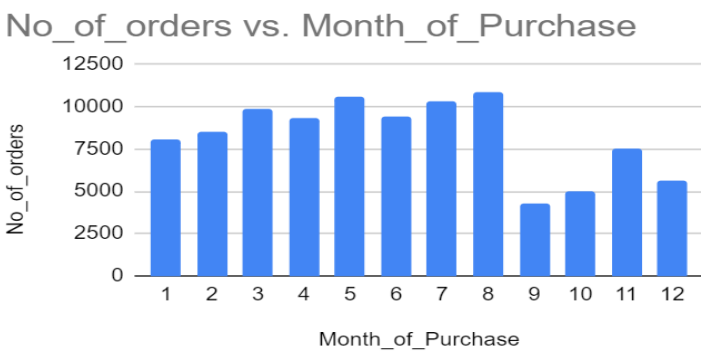


Query: observation 2

```
SELECT * FROM(
SELECT count(order_id) No_of_orders, Extract(MONTH FROM  order_purchase_timestamp) as Month_of_Purchase
from Target.orders
GROUP BY Extract(MONTH FROM order_purchase_timestamp))t
ORDER BY Month_of_Purchase
```

Result: **The maximum orders are placed in the mid of the year from May to August. The highest orders are placed in the month of August.**

No_of_orders	Month_of_Purchase
8069	1
8508	2
9893	3
9343	4
10573	5
9412	6
10318	7
10843	8
4305	9
4959	10
7544	11
5674	12



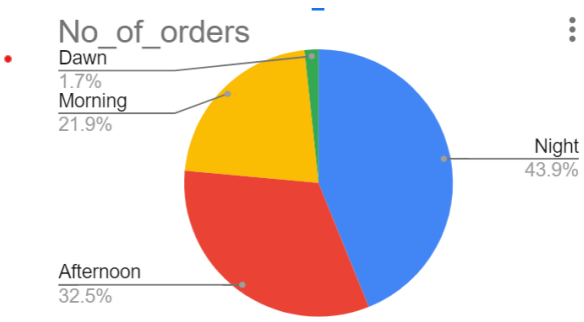
2. **What time do Brazilian customers tend to buy (Dawn, Morning, Afternoon or Night)?**

Query:

```
SELECT COUNT(order_id) as No_of_orders, purchase_hour

 FROM( SELECT order_id,

        case

             WHEN order_hour >1 and order_hour < 7 THEN 'Dawn'

             WHEN order_hour >6 and order_hour < 12 THEN 'Morning'

             WHEN order_hour >12 and order_hour < 18 THEN 'Afternoon'

             ELSE 'Night'

        END as purchase_hour

       FROM (  select order_id, EXTRACT(HOUR from order_purchase_timestamp) as order_hour,
order_purchase_timestamp  from Target.orders ORDER BY order_purchase_timestamp)t1

)t2

GROUP BY purchase_hour

ORDER BY No_of_orders DESC;
```

Result: **Most orders are placed in the night time between 6:00 pm to 2:00 am**

No_of_orders	purchase_hour
43659	Night
32366	Afternoon
21738	Morning
1678	Dawn



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

1.Get month on month orders by region, states

Query:

```
select Extract(YEAR FROM o.order_purchase_timestamp) as Year, Extract(MONTH FROM
o.order_purchase_timestamp) as Month, count(o.order_id) as orders, c.customer_state as
state from Target.orders as o join Target.customers as c

    ON o.customer_id = c.customer_id

    group by Year, Month, state

    ORDER BY Year, Month, state;
```

Result: Only a few columns as a sample for month on month orders based on state.

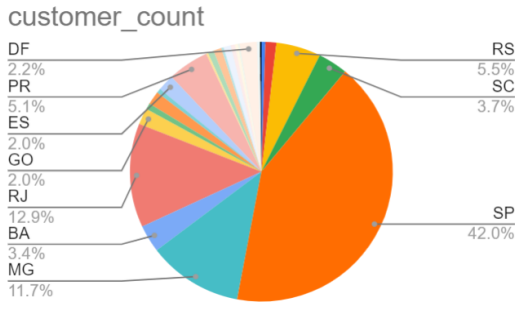
Year	Month	orders	state
2016	9	1	RR
2016	9	1	RS
2016	9	2	SP
2016	10	2	AL
2016	10	4	BA
2016	10	8	CE
2016	10	6	DF
2016	10	4	ES

2.How are customers distributed in Brazil

Query:

```
select c.customer_state as state, count(c.customer_id) as customer_count
from Target.customers as c join Target.orders as o
ON c.customer_id = o.customer_id
group by state
```

Result: **Below is the customer distribution chart for Brazil state vise, Target have more customer from Sao Paulo(SP)**



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)

Query:

```
SELECT ROUND(((SUM(_2018)-SUM(_2017))/SUM(_2017)),2) * 100 as increase_percentage from(

SELECT oi.order_id,(oi.price + oi.freight_value) Total_cost,

EXTRACT(Year from order_purchase_timestamp) as Year

from `Target.order_items` as oi join Target.orders as o

ON oi.order_id = o.order_id

WHERE EXTRACT(Year from order_purchase_timestamp) in (2018,2017) AND EXTRACT(Month from
order_purchase_timestamp) between 1 and 6 )

PIVOT(SUM(Total_cost) FOR Year IN(2018,2017) )
```

Result: *The Approximate percentage increase in the total revenue including freight cost is 179%*

increase_percent...
179.0

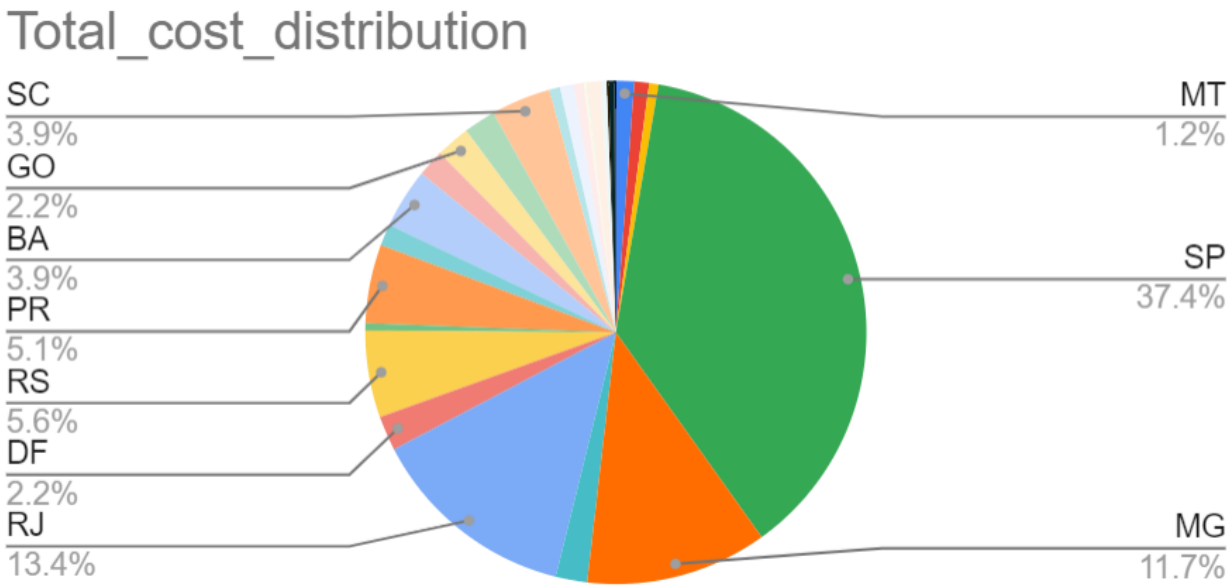
2.Mean & Sum of price and freight value by customer state

Query:

```
SELECT  c.customer_state,AVG(oi.price) as Mean_Price,AVG(oi.freight_value) as Mean_freight_value,

SUM(oi.price) as Total_order_Price,SUM(oi.freight_value) as Total_freight_value ,(SUM(oi.price) +
SUM(oi.freight_value)) Total_cost

FROM Target.customers c join Target.orders as o

ON c.customer_id = o.customer_id

JOIN Target.order_items oi

ON o.order_id = oi.order_id

GROUP BY c.customer_state
```

Result: (showing only 5 rows)

customer_state	Mean_Price	Mean_freigh...	Total_order_...	Total_freigh...	Total_cost
MT	148.297184...	28.1662843...	156453.529...	29715.4300...	186168.959...
MA	145.204150...	38.2570024...	119648.219...	31523.7700...	151171.989...
AL	180.889211...	35.8436711...	80314.81	15914.5899...	96229.4
SP	109.653629...	15.1472753...	5202955.05...	718723.069...	5921678.12...
MG	120.748574...	20.6301668...	1585308.02...	270853.460...	1856161.48...
PE	145.508322...	32.9178626...	262788.029...	59449.6599...	322237.689...
RJ	125.117818...	20.9609239...	1824092.66...	305589.310...	2129681.97...



5.Analysis on sales, freight and delivery time

a. **Calculate days between purchasing, delivering and estimated delivery**
b. **Create columns:**
    i. **time_to_delivery = order_purchase_timestamp-order_delivered_customer_date**
    ii. **diff_estimated_delivery = order_estimated_delivery_date-order_delivered_customer_date**
c. **Group data by state, take mean of freight_value, time_to_delivery, diff_estimated_delivery**

d. **Sort the data to get the following**

Top 5 states with highest average freight value

Query:

```
SELECT customer_state,AVG(freight_value) as
freight_value,AVG(DATE_DIFF(order_delivered_date,order_purchase_date,DAY) ) as
time_to_delivery_inDays,

 AVG(DATE_DIFF(Estimated_delivery_date,order_delivered_date,DAY))as
diff_estimated_delivery

FROM(

SELECT c.customer_state,o.order_id,oi.freight_value, o.order_status,Extract(DATE FROM
o.order_purchase_timestamp) as order_purchase_date ,Extract(DATE FROM
o.order_delivered_customer_date) as order_delivered_date, Extract(DATE FROM
o.order_estimated_delivery_date) as Estimated_delivery_date

FROM Target.orders o JOIN Target.customers c
ON o.customer_id = c.customer_id
JOIN Target.order_items oi
ON o.order_id = oi.order_id
WHERE order_status ='delivered')
GROUP BY customer_state
ORDER BY freight_value DESC
LIMIT 5
```

Result:

customer_state	freight_value
PB	43.0916894...
RR	43.0880434...
RO	41.3305494...
AC	40.0479120...
PI	39.1150860...

Top 5 states with lowest average freight value

Query:

```
SELECT customer_state,AVG(freight_value) as
freight_value,AVG(DATE_DIFF(order_delivered_date,order_purchase_date,DAY) ) as
time_to_delivery_inDays,

 AVG(DATE_DIFF(Estimated_delivery_date,order_delivered_date,DAY))as
diff_estimated_delivery

FROM(

SELECT c.customer_state,o.order_id,oi.freight_value, o.order_status,Extract(DATE FROM
o.order_purchase_timestamp) as order_purchase_date ,Extract(DATE FROM
o.order_delivered_customer_date) as order_delivered_date, Extract(DATE FROM
o.order_estimated_delivery_date) as Estimated_delivery_date

FROM Target.orders o JOIN Target.customers c
ON o.customer_id = c.customer_id
JOIN Target.order_items oi
ON o.order_id = oi.order_id
WHERE order_status ='delivered')
GROUP BY customer_state
ORDER BY freight_value
LIMIT 5
```

Result:

Row	customer_state	freight_value
1	SP	15.1151823...
2	PR	20.4718162...
3	MG	20.6263425...
4	RJ	20.9114360...
5	DF	21.0721613...

Top 5 states with highest average time to delivery

Query:

```
SELECT customer_state,AVG(DATE_DIFF(order_delivered_date,order_purchase_date,DAY) ) as
time_to_delivery_inDays,

 AVG(DATE_DIFF(Estimated_delivery_date,order_delivered_date,DAY))as
diff_estimated_delivery , AVG(freight_value) as freight_value

FROM(

SELECT c.customer_state,o.order_id,oi.freight_value, o.order_status,Extract(DATE FROM
o.order_purchase_timestamp) as order_purchase_date ,Extract(DATE FROM
o.order_delivered_customer_date) as order_delivered_date, Extract(DATE FROM
o.order_estimated_delivery_date) as Estimated_delivery_date

FROM Target.orders o JOIN Target.customers c
ON o.customer_id = c.customer_id
JOIN Target.order_items oi
ON o.order_id = oi.order_id
WHERE order_status ='delivered')
GROUP BY customer_state
ORDER BY time_to_delivery_inDays desc
LIMIT 5
```

Result:

customer_state	time_to_delivery_inDays
AP	28.222222222222218
RR	28.173913043478258
AM	26.337423312883427
AL	24.447306791569098
PA	23.702087286527469

TOP 5 states with lowest average time to delivery

Query:

```
SELECT customer_state,AVG(DATE_DIFF(order_delivered_date,order_purchase_date,DAY) ) as
time_to_delivery_inDays,

 AVG(DATE_DIFF(Estimated_delivery_date,order_delivered_date,DAY))as
diff_estimated_delivery , AVG(freight_value) as freight_value

FROM(

SELECT c.customer_state,o.order_id,oi.freight_value, o.order_status,Extract(DATE FROM
o.order_purchase_timestamp) as order_purchase_date ,Extract(DATE FROM
o.order_delivered_customer_date) as order_delivered_date, Extract(DATE FROM
o.order_estimated_delivery_date) as Estimated_delivery_date

FROM Target.orders o JOIN Target.customers c
ON o.customer_id = c.customer_id
JOIN Target.order_items oi
ON o.order_id = oi.order_id
WHERE order_status ='delivered')
GROUP BY customer_state
ORDER BY time_to_delivery_inDays
LIMIT 5
```

Result:

Row	customer_state	time_to_deli...
1	SP	8.66232423...
2	PR	11.8930784...
3	MG	11.9193248...
4	DF	12.8938428...
5	SC	14.9463021...

TOP 5 states with fastest delivery than the estimated date

Query:

```
SELECT customer_state,
AVG(DATE_DIFF(Estimated_delivery_date,order_delivered_date,DAY))as
diff_estimated_delivery,

AVG(DATE_DIFF(order_delivered_date,order_purchase_date,DAY) ) as
time_to_delivery_inDays,

 AVG(freight_value) as freight_value

FROM(

SELECT c.customer_state,o.order_id,oi.freight_value, o.order_status,Extract(DATE FROM
o.order_purchase_timestamp) as order_purchase_date ,Extract(DATE FROM
o.order_delivered_customer_date) as order_delivered_date, Extract(DATE FROM
o.order_estimated_delivery_date) as Estimated_delivery_date

FROM Target.orders o JOIN Target.customers c
ON o.customer_id = c.customer_id
JOIN Target.order_items oi
ON o.order_id = oi.order_id
WHERE order_status ='delivered')
GROUP BY customer_state
ORDER BY diff_estimated_delivery desc
LIMIT 5
```

Result:

customer_state	diff_estimat...
AC	20.9780219...
RO	20.0402930...
AM	19.9325153...
AP	18.3950617...
RR	18.3260869...

TOP 5 states with slowest delivery than the estimated date

Query:

```
SELECT customer_state,
AVG(DATE_DIFF(Estimated_delivery_date,order_delivered_date,DAY))as
diff_estimated_delivery,

AVG(DATE_DIFF(order_delivered_date,order_purchase_date,DAY) ) as
time_to_delivery_inDays,

 AVG(freight_value) as freight_value

FROM(

SELECT c.customer_state,o.order_id,oi.freight_value, o.order_status,Extract(DATE FROM
o.order_purchase_timestamp) as order_purchase_date ,Extract(DATE FROM
o.order_delivered_customer_date) as order_delivered_date, Extract(DATE FROM
o.order_estimated_delivery_date) as Estimated_delivery_date

FROM Target.orders o JOIN Target.customers c
ON o.customer_id = c.customer_id
JOIN Target.order_items oi
ON o.order_id = oi.order_id
WHERE order_status ='delivered')
GROUP BY customer_state
ORDER BY diff_estimated_delivery
LIMIT 5
```

Result:

customer_state	diff_estimat...
AL	8.73536299...
MA	9.90624999...
SE	10.0026666...
ES	10.6462921...
BA	10.9826228...

6. Payment type analysis:

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

Query:

```
SELECT Extract(MONTH FROM  o.order_purchase_timestamp) as Month,
count(o.order_id) as order_count, p.payment_type

 FROM Target.orders o JOIN Target.payments p

 ON o.order_id = p.order_id

 GROUP BY Month, payment_type

 ORDER BY Month
```

Result: (Showing only for two months)

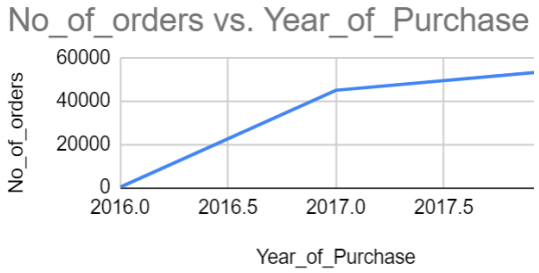
Month	order_count	payment_type
1	477	voucher
1	6103	credit_card
1	118	debit_card
1	1715	UPI
2	6609	credit_card
2	424	voucher
2	1723	UPI
2	82	debit_card

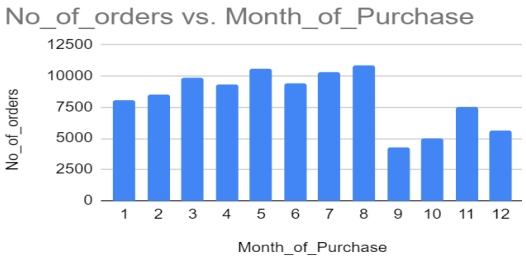
2.Distribution of payment installments and count of orders

```
SELECT  o.order_id as order_count, p.payment_type,p.payment_installments

FROM Target.orders o JOIN Target.payments p

ON o.order_id = p.order_id
```

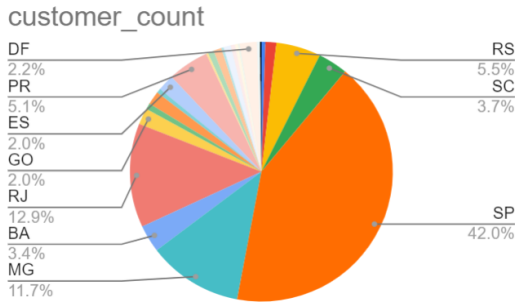
Actionable Insights:

We can see the number of orders placed by the customers increased rapidly over years. Customers started using E-Shopping. The maximum orders are placed in the middle of the year from May to August. The highest orders are placed in the month of August.

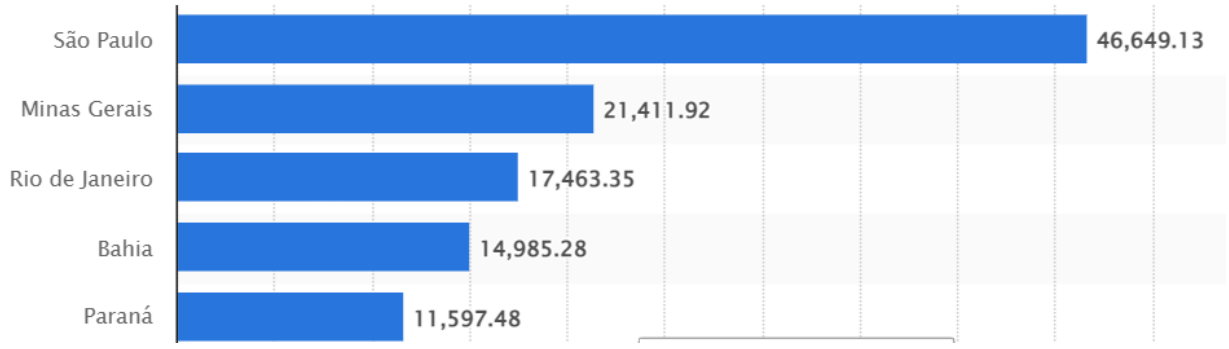




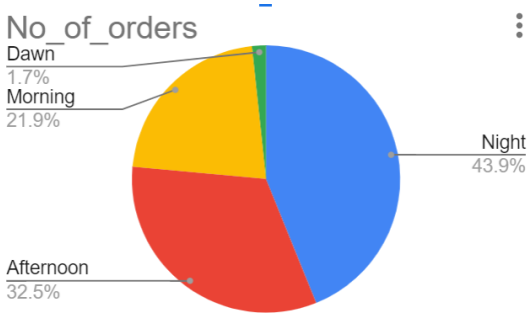
As per the Analyze with the customer count distribution by month most of the customers are from Sao Paulo around 42% of the customer. Advertisement and Campaigns need to be carry out in other highly populated states like Minas, Rio de Janeiro



Brazil Population State wise:



Most of the Orders are placed in late evening and in the night time between 6pm to 2am. So we can run social media ads representing the offers and specials.



Recommendations:

- **Create targeted marketing campaigns and online shout outs which resonate better with your audience by combining actionable insights from both demographic data (age, gender etc.) and consumer behavior,We can also make use of the social media ads to advertise items.**
- **Identify the proper items to sell in specific areas based on the requirements and needs of a particular state. You can also introduce new trends**
- **In some cities freight value is very high. You can have more local distribution partners to reduce the freight cost And we can reduce the delivery time period also. preferred suppliers that you should engage to reduce costs and work (act) with suppliers on discounts etc.**
- **The customer overall rating is 4.0 on average. Can get the feedback from the customers to increase the customer delight.**
- **The number of customers is high in Sao Paulo. We can introduce or test new projects in states like Sao Paulo or Minas.**
- **Also You can run campaigns about online frauds and how to avoid them. It will help to avoid the fear in most people.**