Operational Analysis for a Leading Retailer using BigQuery

1. Exploratory analysis

a. Time period for which the data is given

```
SELECT
     EXTRACT(DATE FROM MIN(order_purchase_timestamp)) MIN_DATE,
     EXTRACT(DATE FROM MAX(order_purchase_timestamp) ) MAX_DATE
FROM `orders`
```

Row	MIN_DATE	MAX_DATE
1	2016-09-04	2018-10-17

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

```
SELECT
  COUNT(DISTINCT customer_city) `Count of Cities`,
  COUNT(DISTINCT customer_state) `Count of States`
FROM `customers`
```

JOB INFORMATION | RESULTS | CHART

Row	Count of Cities	Count of States
1	4119	27

c. Count the Cities & States of sellers.

```
SELECT
  COUNT(DISTINCT seller_city) `Count of Cities`,
  COUNT(DISTINCT seller_state) `Count of States`
FROM `sellers`
```

JOB INFORMATION | RESULTS | CHART

Row	Count of Cities	Count of States
1	611	23

2. Distribution of Customers and Sellers Across States

a. Top 5 States Having Highest Number of Sellers

```
SELECT
  seller_state,
  COUNT(DISTINCT seller_id) `States having highest number of Sellers`
FROM `sellers`
GROUP BY seller_state
ORDER BY 2 DESC
```

Row	seller_state	States having highest number of Sellers
1	SP	1849
2	PR	349
3	MG	244
4	SC	190
5	RJ	171

b. Top 5 States Having Highest Number of Customers

```
SELECT
  customer_state,
  COUNT(DISTINCT customer_id) `States having highest number of Customers`
FROM `customers`
GROUP BY customer_state
ORDER BY 2 DESC
```

Row	customer_state	States having highest number of Customers
1	SP	41746
2	RJ	12852
3	MG	11635
4	RS	5466
5	PR	5045

Observations:

1) São Paulo(SP) has the highest number of customers and sellers.
2) Considering the above results Parana(PR) has more Sellers than customers and Rio De Janeiro(RJ) has the opposite.

More insights could be derived if we can find out states with higher customer/seller ratio.

3. Trends and Seasonality

a. Is there a growing trend on e-commerce for this dataset? Can we see some seasonality with peaks at specific months?

```
SELECT
  Year, Quarter, ROUND(Sales,2) Sales,ROUND(c_sales,2) `Cumulative Sales`,
  ROUND((Sales-LAG(sales) OVER(ORDER BY year,quarter))*100/LAG(sales) OVER(ORDER
BY year,quarter),2) `QoQ %age Growth`
FROM
(
  SELECT DISTINCT
    year,quarter,
    SUM(sales) OVER(PARTITION BY year,quarter) Sales,
    SUM(sales) OVER(ORDER BY year,quarter) c_sales,
  FROM
  (
    SELECT
      year,
      CASE WHEN month BETWEEN 1 AND 3 THEN 'Q1'
      WHEN month BETWEEN 4 AND 6 THEN 'Q2'
      WHEN month BETWEEN 7 AND 9 THEN 'Q3' ELSE 'Q4' END quarter, sales
```

```
    FROM
    (
      SELECT
        EXTRACT(YEAR FROM o.order_purchase_timestamp) year,
        EXTRACT(MONTH FROM o.order_purchase_timestamp) month,
        p.payment_value
      FROM `orders` o, `payments` p
      WHERE o.order_id = p.order_id
    )
  )
)
ORDER BY 1,2
```

Row	Year	Quarter	Sales	Cumulative Sales	QoQ %age Growth
1	2016	Q3	252.24	252.24	null
2	2016	Q4	59110.1	59362.34	23334.07
3	2017	Q1	880259.65	939621.99	1389.19
4	2017	Q2	1521983.23	2461605.22	72.9
5	2017	Q3	1994541.69	4456146.91	31.05
6	2017	Q4	2852962.16	7309109.07	43.04
7	2018	Q1	3267119.64	10576228.71	14.52
8	2018	Q2	3338648.13	13914876.84	2.19
9	2018	Q3	2093405.61	16008282.45	-37.3
10	2018	Q4	589.67	16008872.12	-99.97

```
SELECT
  Year, Month, orders `No of orders`, ROUND(Sales,2) Sales, ROUND(c_sales,2)
`Cumulative Sales`,
  ROUND((Sales-LAG(sales) OVER(ORDER BY year,month))*100/LAG(sales) OVER(ORDER BY
year,month),2) `MoM %age Growth`,
FROM
(
  SELECT DISTINCT
    year,month,
    SUM(payment_value) OVER(PARTITION BY year,month) Sales,
```

```
    SUM(payment_value) OVER(ORDER BY year,month) c_sales,
    COUNT(order_id) OVER(PARTITION BY year,month) orders
  FROM
  (
    SELECT
        EXTRACT(YEAR FROM o.order_purchase_timestamp) year,
        EXTRACT(MONTH FROM o.order_purchase_timestamp) month,
        p.payment_value, o.order_id
    FROM `orders` o, `payments` p
    WHERE o.order_id = p.order_id

  )
)
ORDER BY 1,2
```

Row	Year	Month	No of orders	Sales	Cumulative Sales	MoM %age Growth
1	2016	9	3	252.24	252.24	*null*
2	2016	10	342	59090.48	59342.72	23326.29
3	2016	12	1	19.62	59362.34	-99.97
4	2017	1	850	138488.04	197850.38	705751.38
5	2017	2	1886	291908.01	489758.39	110.78
6	2017	3	2837	449863.6	939621.99	54.11
7	2017	4	2571	417788.03	1357410.02	-7.13
8	2017	5	3944	592918.82	1950328.84	41.92
9	2017	6	3436	511276.38	2461605.22	-13.77
10	2017	7	4317	592382.92	3053988.14	15.86

Observations:

1) Yes, e-commerce is overall growing.
2) At the start of every year(January) the sales and number of orders goes up. As this dataset is of Brazil. This spike might be due to the Rio Carnival which is held in the First quarter of every year.

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

```sql
SELECT phases, COUNT(order_id) number_of_orders, ROUND(SUM(sales),2) total_sales
FROM
(
      SELECT
            CASE WHEN hour BETWEEN 3 AND 5 THEN 'Dawn'
            WHEN hour BETWEEN 6 AND 11 THEN 'Morning'
            WHEN hour BETWEEN 12 AND 16 THEN 'Afternoon'
            WHEN hour BETWEEN 17 AND 20 ELSE 'Night' END phases,
            sales, order_id
      FROM
      (
            SELECT
                  EXTRACT(HOUR FROM o.order_purchase_timestamp) hour,
                  p.payment_value sales, o.order_id
            FROM `orders` o, `payments` p
            WHERE o.order_id = p.order_id
      )
)
GROUP BY phases
```

Row	phases	number_of_orders	total_sales
1	Afternoon	33603	5299071.83
2	Evening	25152	3925741.3
3	Morning	23132	3541310.97
4	Night	21308	3146034.54
5	Dawn	691	96713.48

Observations:

1) Brazilian Customers tend to buy mostly at Afternoon and Evening

4. Impact of Money

a. **Avg & Total of price and freight value by customer state**

```
SELECT
  c.customer_state State, ROUND(AVG(oi.price),2) `Avg Price`,
  ROUND(SUM(oi.price),2) `Total Price`,
  ROUND(SUM(p.payment_value),2) `Total Sales`,
  ROUND(AVG(p.payment_value),2) `Avg order Value`,
  ROUND(AVG(oi.freight_value),2) `Avg Freight value`,
  ROUND(SUM(oi.freight_value),2) `Total Freight value`,
FROM
  `orders` o, `order_items` oi, `customers` c, `payments` p
WHERE o.order_Id = oi.order_id AND o.customer_id = c.customer_id AND p.order_id =
o.order_id
GROUP BY c.customer_state
ORDER by 1
```

Row	State	Avg Price	Total Price	Total Sales	Avg order Value	Avg Freight value	Total Freight value
1	PB	193.62	123726.34	180984.19	283.23	43.26	27641.72
2	AC	179.57	17059.44	24984.86	263.0	40.02	3802.06
3	AP	162.55	13654.3	21642.7	257.65	34.08	2863.09
4	AL	181.46	83110.36	111284.42	242.98	35.74	16368.65
5	RR	150.57	7829.43	12462.21	239.66	42.98	2235.19
6	PI	160.85	92167.7	136779.96	238.71	39.23	22480.64
7	MA	145.59	122881.79	198566.27	235.27	38.26	32290.33
8	PA	165.24	184407.88	261788.35	234.58	35.74	39881.05
9	RO	164.21	46964.03	65886.0	230.37	40.97	11717.47
10	MT	151.84	170822.04	256804.62	228.27	28.97	32592.32

b. **Find avg Freight to Order Value Ratio and Price to Order Value Ratio**

```
SELECT
  c.customer_state State,
ROUND((AVG(oi.freight_value)/AVG(p.payment_value))*100,2) `Freight to Order value
Ratio`,
  ROUND((AVG(oi.price)/AVG(p.payment_value))*100,2) `Avg Price to Order Value
Ratio`

FROM
  `orders` o, `order_items` oi, `customers` c, `payments` p
WHERE o.order_Id = oi.order_id AND o.customer_id = c.customer_id AND p.order_id =
o.order_id
```

```
GROUP BY c.customer_state
ORDER by 1
```

Row	State	Freight to Order value Ratio	Avg Price to Order Value Ratio
1	TO	18.61	77.82
2	RR	17.94	62.83
3	RO	17.78	71.28
4	RN	17.27	81.33
5	PE	16.45	72.07
6	PI	16.44	67.38
7	SE	16.44	68.92
8	AM	16.28	65.63
9	MA	16.26	61.88
10	PB	15.27	68.36

Observations:

1) Total Price is lesser than Total Sales, because of taxes, shipping charges, EMI rates.
2) State Sao Paulo(SP) has the highest Total Sales, Total Price of Products and Freight Value.
3) State Paraiba(PB) has the highest avg Price of Products and avg order value per state which is almost twice of avg price and avg order value in Sao Paulo.
4) High avg price and average order value in Pariba, might be due to high operational costs/tourism, tax and regulations levied by the govt or Logistics cost or Premium Users.
5) State Rio Grande do Norte (RN) has the highest avg price to order value ratio, which denotes low taxes are levied in State RN.
6) State Tocantis(TO) has the highest AVG freight to Order Ratio and State Sao Paulo(SP) has the lowest. This indicates lower shipping charges for Sao Paulo.

5. Analysis on sales, freight and delivery time

a. **Calculate days between purchasing, delivering and estimated delivery**

```
SELECT order_id,
 DATE_DIFF(order_delivered_customer_date, order_purchase_timestamp, DAY)
`days taken to deliver`,
 DATE_DIFF(order_estimated_delivery_date, order_delivered_customer_date, DAY)
`diff estimated delivery` ,
FROM `orders`
ORDER BY 1
```

Row	order_id	days taken to deliver	diff estimated delivery
1	00010242fe8c5a6d1ba2dd792...	7	8
2	00018f77f2f0320c557190d7a1...	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

Observations:

1) Around 3% of the order_delivered_customer_date are blank/null, maybe due to some missing records.
2) Including these records around 87.5% of the orders delivered were delivered before the estimated delivery date.

b. **Group data by state, take mean of freight_value, time_to_delivery, diff_estimated_delivery**

```
SELECT c.customer_state State ,
 ROUND(AVG(oi.freight_value),2) `Avg Freight Value`,
ROUND(AVG(DATE_DIFF(order_delivered_customer_date,order_purchase_timestamp,DAY)),
2) `Avg time of purchase to delivery`,
ROUND(AVG(DATE_DIFF(order_estimated_delivery_date,order_delivered_customer_date
,DAY)),2) `Avg estimated time of purchase to delivery`
FROM `orders` o, `order_items` oi, `customers` c
WHERE o.order_id = oi.order_id AND o.customer_id = c.customer_id
GROUP BY c.customer_state
```

Observations:

1) RR, AP, AM, AL, PA are the states who get late deliveries (mean of time to delivery).
2) SP, PR, MG, DF, SC are the states who get the earliest deliveries (mean of time to delivery).
3) RR, PB, RO, AC, PI has the highest amount of freight values. RR freight value is 15% more than the average freight value
4) By c and a we can assume that state RR has some transportation difficulties as its freight values are high as well as time taken to deliver to the state is also high.

6. Payment type analysis

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

```
SELECT
      EXTRACT(YEAR FROM o.order_purchase_timestamp) year,
      EXTRACT(month FROM o.order_purchase_timestamp) month,
      CASE WHEN payment_type = 'UPI' THEN COUNT(p.order_id) ELSE 0 END online,
      CASE WHEN payment_type = 'credit_card' THEN COUNT(p.order_id) ELSE 0 END
`credit card`,
      CASE WHEN payment_type = 'voucher' THEN COUNT(p.order_id) ELSE 0 END
voucher,
      CASE WHEN payment_type = 'debit_card' THEN COUNT(p.order_id) ELSE 0 END
`debit card`,
      CASE WHEN payment_type = 'not_defined' THEN COUNT(p.order_id) ELSE 0 END NA
FROM `payments` p, `orders` o
WHERE o.order_id = p.order_id
GROUP BY EXTRACT(YEAR FROM o.order_purchase_timestamp), EXTRACT(month FROM
o.order_purchase_timestamp), p.payment_type
ORDER BY 1,2
```

Row	year	month	online	credit card	voucher	debit card	NA
9	2017	1	0	583	0	0	0
10	2017	1	0	0	0	9	0
11	2017	2	0	1356	0	0	0
12	2017	2	0	0	119	0	0
13	2017	2	398	0	0	0	0
14	2017	2	0	0	0	13	0
15	2017	3	0	0	200	0	0
16	2017	3	590	0	0	0	0
17	2017	3	0	2016	0	0	0
18	2017	3	0	0	0	31	0
19	2017	4	0	0	202	0	0

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

```
SELECT payment_installments, COUNT(order_id) count_of_orders
FROM `payments`
GROUP BY payment_installments
```

Row	payment_installments	count_of_orders
1	1	52546
2	2	12413
3	3	10461
4	4	7098
5	10	5328
6	5	5239
7	8	4268
8	6	3920
9	7	1626
10	9	644

Observations:

1) 1, 2, 3, 4, 10 are the most preferred payment_installments.
2) Most preferred payment type is credit card then UPI followed by debit card

Final Insights:

- São Paulo(SP) has the highest number of customers and sellers and sales and lower shipping charges.
- State Paraíba (PB) has the highest average order values, making it fit for premium products.
- Average Price to Payment Ratio is highest in the State Rio Grande do Norte (RN), which means lower taxes during billing.
- The state of Roraima (RR) presents logistical challenges, it is evident, due to its higher shipping charges and more time taken for delivery.
- At the start of every year (January) the sales and number of orders goes up. This spike might be due to the Rio Carnival, which is held in the first quarter of every year.
- Brazilian Customers tend to buy mostly at Afternoon and Evening
- 87.5% of the orders delivered were delivered before the estimated delivery date.
- Among the available installment options, those spanning 1, 2, 3, 4, and 10 monthly installments are the most preferred.
- Most preferred payment type is credit card, followed by online and debit card.
- Bed Table Bath is the most bought product category followed by Health Beauty, Sport Leisure, Furniture Decorations and Computer accessories.
- Customers tend to buy the least items from these product categories (Art, PCS, Signalization and Safety, Construction and Tools and Christmas articles).

Recommendations:

- The retailer should consider partnering with a credit card company to promote their card and run advertisements. This strategy could incentivize customers, as the majority of customers, prefer using credit cards for payments.

- Offering attractive schemes like no-cost EMI options for up to 12 installments would cater to customer preferences for convenient payment plans.
- For orders from the state of RR, the retailer should prioritize optimizing logistics operations, as this state has longer delivery times and higher freight costs.
- To capitalize on the market in the state of PB, the retailer could consider recommending and promoting luxurious or premium range products. Customers in this region have a higher average order value.
- For customers in the state of SP, the retailer should explore offering multiple delivery options, such as fast and standard shipping, with corresponding pricing tiers.
- The retailer should consider running promotional sales or campaigns between 1300 and 1800 hours, as data indicates customers tend to make more purchases during these hours.
- To drive sales, the retailer could feature and promote products from categories like Bed Table Bath, Health Beauty, Sport Leisure, Furniture Decorations, and Computer Accessories, as customers tend to purchase more items from these categories.
- For product categories with lower demand, the retailer could implement clearance sales or promotions to clear inventory and stimulate demand.