# 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.

```
SELECT
column_name,
data_type
FROM
`target-403119.target_casestudy.INFORMATION_SCHEMA.COLUMNS`
WHERE
table_name = 'customers';
```

.

**Insights**:

There are 4 coloumn with STRING data type and one with INTEGER.\

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

```
select

min(order_purchase_timestamp) as start_date ,

max(order_purchase_timestamp) as end_date

From `target_casestudy.orders`
```

**Insights:**

- with this query we can check that the data is from 2016 September to 2018 October
- The first order was on 4th September

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

```
select count(distinct customer_city) as No_of_cities ,

count(distinct customer_state) as No_of_states

from `target_casestudy.customers`

where customer_id in (select customer_id from `target_casestudy.orders`)
```

**Insights**:

- There are 4119 different cities and 27 different states from where the company got the orders between 2016 till 2018.

**Recommendations:**

- Company can open the warehouses at least in the metro cities of these states to reduce the delivery time of the order which will fetch more customer satisfaction and ultimately repeat orders from the same customers.
- Keep more inventory in these regions.

## 2 In-depth Exploration:

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

```
select

extract(year from order_purchase_timestamp) as year ,

extract(month from order_purchase_timestamp) as month,

count(order_id) as no_of_orders

From `target_casestudy.orders`

group by 1 ,2

order by 1 , 2
```

| | | | |
|---|---|---|---|
| 12 | 2017 | 9 | 4285 |

**Insights**:

- It is visible that we got a spike on number of orders from 2017 Jan till 2017 Nov. Post that it decreased a lil bit and then started growing again till august 2018 then suddenly it drastically fallen.
- Also, in 2016 September-October and 2018 September-October their sales were almost negligible.

**Recommendations**:

- Run lucrative offer during September- October time to attract the customers

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

```
select *,
round(((lead(no_of_orders,1)over(order by year, month)-
 no_of_orders)/no_of_orders)*100,0)as percent_hike
 from
(
select
extract(year from order_purchase_timestamp) as year ,
extract(month from order_purchase_timestamp) as month,
count(order_id) as no_of_orders
from `target_casestudy.orders`
group by 1 ,2
order by 1 , 2 ) temp
order by year, month
```

| | | | | |
|---|---|---|---|---|
| 10 | 2017 | 7 | 4026 | 8.0 |

**Insights:**

- **T**he data fluctuates every month but not in some fashion.
- It is visible that we got a spike on number of orders from 2017 Jan till 2017 Nov.
- Post that it decreased a lil bit and then started growing again till august 2018 then suddenly it drastically fallen. Also, in 2016 September-October and 2018 September-October their sales were almost negligible.

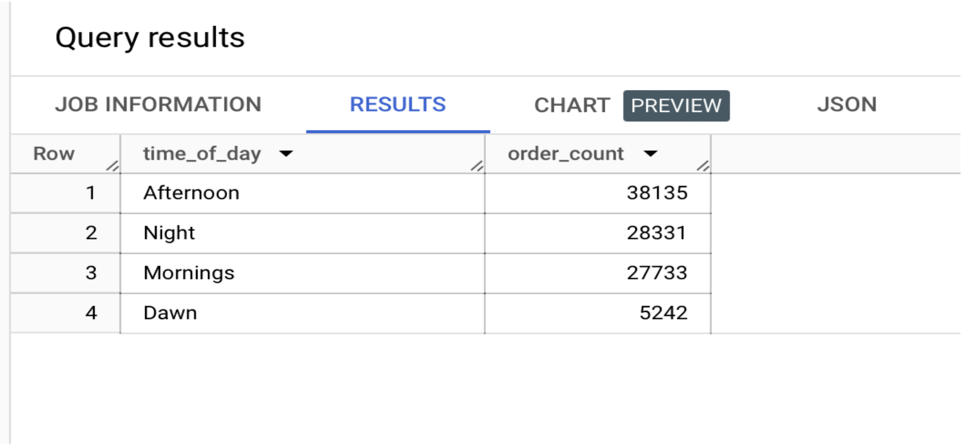
**Recommendation:**

- Run lucrative deals and buy one get one offers to attract customers also run advertising campaign to get the visibility

6. 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

```
select time_of_day ,
count(*) as order_count
from
(
select order_purchase_timestamp,
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
'Mornings'
 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'
 end as time_of_day
 from target_casestudy.orders ) temp
 group by 1
 order by 2 desc
```



Query results

| Row | time_of_day | order_count |
|---|---|---|
| 1 | Afternoon | 38135 |
| 2 | Night | 28331 |
| 3 | Mornings | 27733 |
| 4 | Dawn | 5242 |

**Insights:**

- We can see the highest no. Orders are during the afternoon time, 2nd highest is at nighttime, 3rd is at morning. We can clearly see the traffic is huge during the afternoon, then at night and then morning.

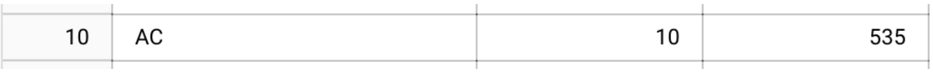
**Recommendation:**

- can run lucrative deals during these peak hours to get the spike in orders also ads can be run of similar products of what the customer is searching or buying to get the visibility in other products plus to get order of those advertised product as well.
- Similarly, we can make buy one get one offer also a package of 2,3,4 etc kind of deals.
- During dawn time to get traffic we can send customized mail and messages of multiple lucrative offers as mentioned above. Plus, we can run some timely deals like 2-hour, 1 hour kind of offers on app to increase the traffic during dawn time.

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

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

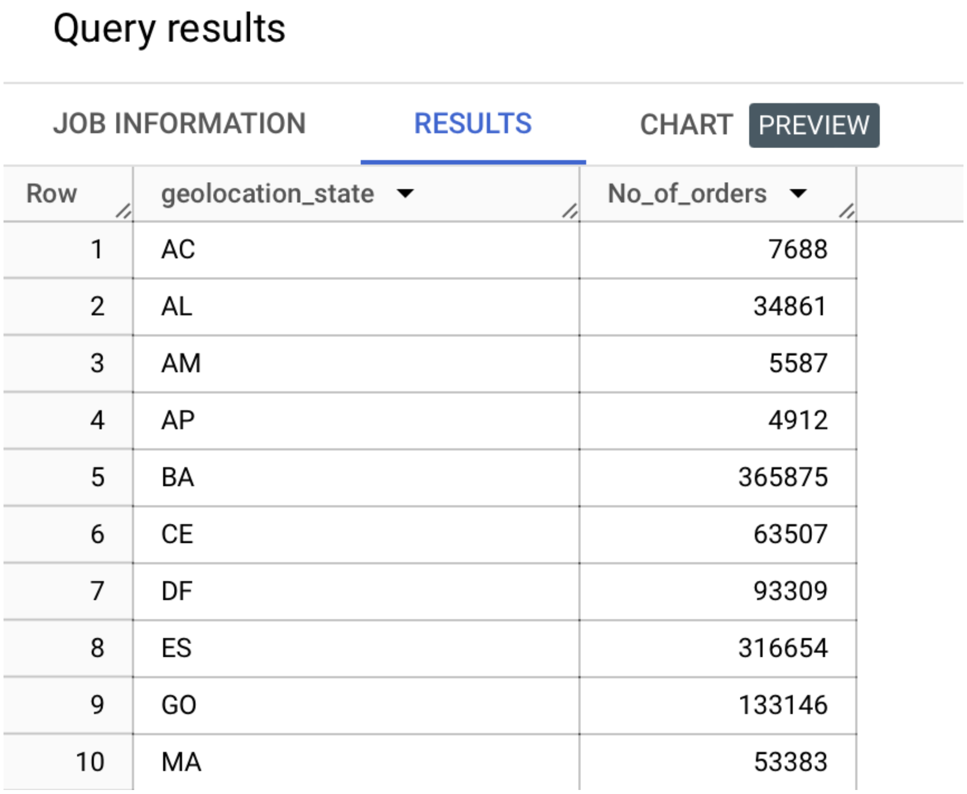
```
 SELECT geolocation_state ,
 extract(month from order_purchase_timestamp) as Months,
 COUNT(*) as No_of_orders
 FROM `target_casestudy.orders` O
 LEFT JOIN `target_casestudy.customers` C ON O.customer_id = C.customer_id
 JOIN `target_casestudy.geolocation` G ON C.customer_zip_code_prefix =
 G.geolocation_zip_code_prefix
GROUP BY geolocation_state , Months
ORDER BY geolocation_state, Months;
```



| | | | |
|---|---|---|---|
| 10 | AC | 10 | 535 |

```
SELECT geolocation_state ,
COUNT (*) as No_of_orders
FROM `target_casestudy.orders` O
LEFT JOIN `target_casestudy.customers` C ON O.customer_id = C.customer_id
```

```
JOIN `target_casestudy.geolocation` G ON C.customer_zip_code_prefix =
G.geolocation_zip_code_prefix
GROUP BY geolocation_state
ORDER BY geolocation_state
```



| Row | geolocation_state | No_of_orders |
|---|---|---|
| 1 | AC | 7688 |
| 2 | AL | 34861 |
| 3 | AM | 5587 |
| 4 | AP | 4912 |
| 5 | BA | 365875 |
| 6 | CE | 63507 |
| 7 | DF | 93309 |
| 8 | ES | 316654 |
| 9 | GO | 133146 |
| 10 | MA | 53383 |

**Insights**:

- By this we understood that we are getting orders from 27 different states. For state AC we can see that the no. Of orders are higher during May and August similarly we can check for each state. It can help identify whether certain states have different ordering behaviors or preferences. For example, we might find that some states tend to order more in the summer, while others prefer the winter.
- Also, the highest no. Of orders that the company received is from RJ and second highest is MG.
- By analyzing the month-on-month order counts, we can identify seasonal variations in different states. For example, in some states, we might observe higher order counts during holiday seasons or specific times of the year.
- It can help identify whether certain states have different ordering behaviors or preferences. For example, we might find that some states tend to order more in the summer, while others prefer the winter

**Recommendations:**

- The geographical distribution shows us the order distribution across different states which will help in optimizing the marketing campaigns to target the specific region as well as help in operational efforts. The data can be used for planning inventory, staffing, or logistics. Knowing when and where orders are likely to peak can help businesses manage their resources more effectively.

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

```
 SELECT
geolocation_state,
COUNT(DISTINCT customer_id) as Customer_Count
FROM `target_casestudy.customers` C
JOIN `target_casestudy.geolocation` G ON C.customer_zip_code_prefix =
G.geolocation_zip_code_prefix
GROUP BY geolocation_state
ORDER BY Customer_Count DESC;
```

**Insights:**

- Major base of customers are in SP, RJ, MG. Lowest are in AC, AP, RR.

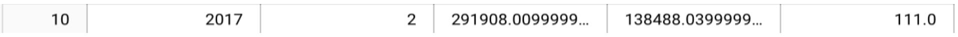
**Recommendations:**

- In regions with a high customer count, it's essential to focus on customer retention and satisfaction to maintain and potentially grow your customer base further. We can have customer service team here to take the feedback from the customers
- Local stores can be opened in low customer concentrated regions to create visibility about the brand.
- More distribution centers can be opened in more customer concentrated region for the quick delivery. can optimize supply chain and delivery routes based on customer locations.
- can tailor marketing and product offerings to meet the needs and preferences of customers in specific states.

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

9. Get the % increase in the cost of orders from 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.

```
      with final_table as
   (
   select
   extract (year from order_purchase_timestamp) as order_year ,
   extract (month from order_purchase_timestamp) as order_month ,
   sum(payment_value) as cost ,
   from `target_casestudy.orders` O
   inner join `target_casestudy.payments` P
   on O.order_id = P.order_id
   where extract (year from order_purchase_timestamp) = 2017 or extract (year from
      order_purchase_timestamp) = 2018 and
   extract (month from order_purchase_timestamp) between 1 and 8
   group by 1 , 2
   )
   select order_year , order_month , cost ,
   LAG(cost, 1) OVER (PARTITION BY order_year ORDER BY order_month) AS
    prev_year_cost,
   round(((cost- LAG(cost, 1) OVER (PARTITION BY order_year ORDER BY order_month)) /
    LAG(cost, 1) OVER (PARTITION BY order_year ORDER BY order_month))*100 ,0) as
    percent_increase
   from final_table
   order by 1 desc , 2
```



| | | | | | |
|---|---|---|---|---|---|
| 10 | 2017 | 2 | 291908.0099999... | 138488.0399999... | 111.0 |

Approach 2 for yearly percentage

```
WITH final_table AS (
SELECT
EXTRACT(YEAR FROM order_purchase_timestamp) AS order_year,
EXTRACT(MONTH FROM order_purchase_timestamp) AS order_month,
SUM(payment_value) AS cost
FROM
target_casestudy.orders O
INNER JOIN target_casestudy.payments P ON O.order_id = P.order_id
WHERE
(EXTRACT(YEAR FROM order_purchase_timestamp) = 2017 OR EXTRACT(YEAR FROM order_purchase_timestamp) = 2018)
AND EXTRACT(MONTH FROM order_purchase_timestamp) BETWEEN 1 AND 8
GROUP BY 1, 2
)
SELECT
order_year,
SUM(cost) AS total_cost,
LAG(SUM(cost), 1) OVER (ORDER BY order_year) AS prev_year_cost,
ROUND(((SUM(cost) - LAG(SUM(cost), 1) OVER (ORDER BY order_year)) / LAG(SUM(cost), 1) OVER (ORDER BY order_year)) * 100, 0) AS percent_increase
FROM final_table
GROUP BY order_year
ORDER BY order_year;
```

**Insights:**

- There is increase of 137% in total cost from 2017 to 2018

**Recommendations :**

- To increase the sales further the company needs to expand in more states and more remote areas. If the visibility increases, then the sales will increase

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

```
SELECT geolocation_state ,
sum(payment_value) as Total ,
avg(payment_value) as Average
from `target_casestudy.payments` P
inner join `target_casestudy.orders` O
on P.order_id = O.order_id
inner join `target_casestudy.customers` C
on O.customer_id= C.customer_id
inner join `target_casestudy.geolocation` G
on C.customer_zip_code_prefix = G.geolocation_zip_code_prefix
```

```
group by 1

order by 2 desc, 3 desc
```

| | | | |
|---|---|---|---|
| 10 | GO | 24572389.12999... | 178.1040710582... |

**Insights:**

- The highest revenue is generated from SP, RJ and MG state. Similarly, we can check the average value of order from each state, here the highest is from the PB, AL, AP, AC, AI, state.

**Recommendations:**

- Focus on high value states
- States with higher average order values, such as "AL", "PB", and "CE" may present opportunities for targeting high-value customers. Consider marketing initiatives or promotions that specifically target these states to further increase sales and revenue.
- Analyze low value states and understand the reasons behind the low value one of the reasons could be economical factors, address these factors will boost up the sales as we can customized and create targeted marketing campaigns and promotions for each segment.

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

```
select geolocation_state ,
sum(freight_value) as Total_freight_value,
avg(freight_value) as Average_freight_value
from `target_casestudy.order_items` OI
inner join `target_casestudy.sellers` S
on OI.seller_id = S.seller_id
inner join `target_casestudy.geolocation` G
on S.seller_zip_code_prefix = G.geolocation_zip_code_prefix
group by 1
order by 2 desc , 3 desc
```

**Insights:**

- There is a noticeable variation in both the total and average freight values across different states. Eg: SP has the highest total freight value and relatively lower average freight value indicating that high volume of orders with low freight values.
- Similarly higher average freight values indicate that the customers in these states are paying relatively high for shipping.

**Recommendation:**

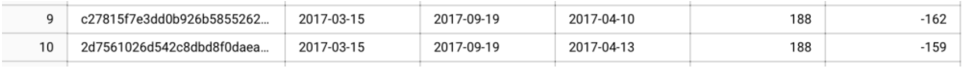
- In the states with the higher average freight value, we can change the logistic partners to reduce the costs. Efficient logistics will help in reducing the shipping cost and ultimately will help in customer satisfaction.

## 5 Analysis based on sales, freight and delivery time.

12. 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.
    Do this in a single query.

    You can calculate the delivery time and the difference between the estimated & actual delivery date using the given formula:
    **time_to_deliver** = order_delivered_customer_date - order_purchase_timestamp
    **diff_estimated_delivery** = order_estimated_delivery_date - order_delivered_customer_date

```
with final_delivery as (
select order_id,
extract(date from order_purchase_timestamp) as purchase_date ,
extract(date from order_delivered_customer_date) as delivery_date ,
extract(date from order_estimated_delivery_date) as estimated_delivery_date
from `target_casestudy.orders` )
select * ,
date_diff(delivery_date , purchase_date , day) as delivery_time ,
date_diff(estimated_delivery_date , delivery_date , day) as difference
from final_delivery
order by delivery_time desc
```



| | | | | | | |
|---|---|---|---|---|---|---|
| 9 | c27815f7e3dd0b926b5855262... | 2017-03-15 | 2017-09-19 | 2017-04-10 | 188 | -162 |
| 10 | 2d7561026d542c8dbd8f0daea... | 2017-03-15 | 2017-09-19 | 2017-04-13 | 188 | -159 |

**Insights:**

- The "difference" column indicates the difference between the estimated delivery date and the actual delivery date.
- Negative values in this column indicate that the orders were delivered before the estimated date, while positive values indicate orders that were delivered after the estimated date. The largest positive difference in this dataset is -144 days, indicating that an order was delivered 144 days later than estimated.
- Many orders have a delivery time that exceeds the estimated delivery date by a consistent number of days. This consistency may suggest a systematic issue in the delivery process or a need for more accurate estimates.
- It's important to consider how these delivery times align with customer expectations. Longer delivery times, especially when significantly exceeding the estimated delivery date, can lead to customer dissatisfaction.

**Recommendation:**

- Need to work with the logistics partners to reduce the delivery time also the arrangements and inventory of the fulfillment centers needs to be checked on regular basis to avoid the higher delivery time.
- Need to understand what contributed to the shortest delivery time orders and accordingly need to work on the rest.

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

```
    with average as (
    select geolocation_state ,
    avg(freight_value) as avg_value
    from `target_casestudy.order_items` OI
    inner join `target_casestudy.sellers` S
    on OI.seller_id = S.seller_id
    Inner join `target_casestudy.geolocation` G
    on S.seller_zip_code_prefix = G.geolocation_zip_code_prefix
    group by 1 )

    select A.geolocation_state as Top_state , A.avg_value , B.geolocation_state as
    Bottom_state , B.avg_value
    from
(select geolocation_state , avg_value , row_number()over(order by avg_value desc) as  R1
  from average
  order by 2 desc
  limit 5 ) A
```

```
 join
(select geolocation_state , avg_value , row_number()over(order by avg_value asc) R2
from average
order by 2 asc
 limit 5 ) B
on A.R1 = B.R2
```

**Insights:**

- CE, RO, PI , PB, AC are the top 5 states with the highest average freight value similarly RN, SP, RJ, DF, PR are bottom 5.
- Highest average freight value is around 54.44 in the entire orders data and lowest is around 22.1

**Recommendation:**

- Need to work on supply chain and logistics in the states with higher average freight value.
- For cost optimization we need to work on the supply chain if the volume of orders is not high then the cost of shipping must be high in the region.
- Similarly, states at the bottom can provide opportunities for performance improvement there might me low volume of orders hence the average cost is low, we can optimize the targeted marketing strategies to get more visibility.

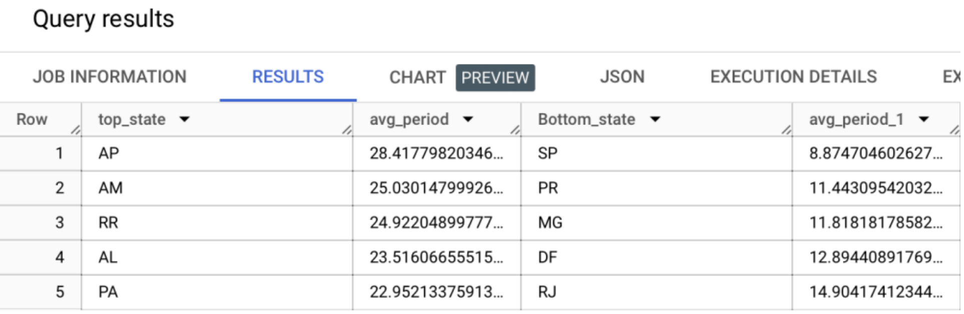
**We got the no. Of orders for each state previously that can be used here.**

```
SELECT geolocation_state ,
COUNT(*) as No_of_orders
FROM `target_casestudy.orders` O
LEFT JOIN `target_casestudy.customers` C ON O.customer_id = C.customer_id
JOIN `target_casestudy.geolocation` G ON C.customer_zip_code_prefix = G.geolocation_zip_code_prefix
where geolocation_state in ('CE', 'RO' ,'PI' , 'PB', 'AC')
GROUP BY geolocation_state
ORDER BY geolocation_state
```

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

```
 with delivery_data as (
select order_id, geolocation_state,
extract(date from order_purchase_timestamp) as purchase_date ,
extract(date from order_delivered_customer_date) as delivery_date ,
date_diff(extract(date from order_delivered_customer_date),extract(date from
     order_purchase_timestamp), day) as period
from `target_casestudy.orders` O
join `target_casestudy.customers` C
on O.customer_id = C.customer_id
join `target_casestudy.geolocation` G
on C.customer_zip_code_prefix = G.geolocation_zip_code_prefix )


 select A.geolocation_state as top_state , A.avg_period , B.geolocation_state as
     Bottom_state , B.avg_period
 from
 (
 select geolocation_state , avg(period) as avg_period ,
 row_number()over(order by avg(period) desc) as R1
 from delivery_data
 group by 1
 order by 2 desc
 limit 5
) A
join
(
select geolocation_state , avg(period) as avg_period ,
row_number()over(order by avg(period) asc ) as R2
from delivery_data
group by 1
order by 2 asc
limit 5 ) B
on A.R1= B.R2
```



| Row | top_state | avg_period | Bottom_state | avg_period_1 |
|---|---|---|---|---|
| 1 | AP | 28.41779820346... | SP | 8.874704602627... |
| 2 | AM | 25.03014799926... | PR | 11.44309542032... |
| 3 | RR | 24.92204899777... | MG | 11.81818178582... |
| 4 | AL | 23.51606655515... | DF | 12.89440891769... |
| 5 | PA | 22.95213375913... | RJ | 14.90417412344... |

**Insights:**

- States with the highest average delivery time are AP, AM, RR, AL, PA. Similarly with lowest delivery time are SP, PR, MG, DF, RJ.
- There is considerable variation in delivery time across the different states.
- The reasons behind these differences may be influenced by factors such as logistics infrastructure, geographical location, and transportation efficiency.
- AM has the second-highest average delivery time at around 25.03 days. Given its geographical location, which includes remote areas of the Amazon rainforest, extended delivery times are not surprising. This highlights the importance of efficient logistics and infrastructure in such regions.
- PA has an average delivery time of roughly 22.95 days. This state, known for its vast rainforest and rivers, may face delivery challenges due to its geography. Businesses operating in Pará should consider strategies for more efficient deliveries.
- SP (São Paulo) stands out with the lowest average delivery time of approximately 8.87 days. This state's highly developed infrastructure and urban centers contribute to faster deliveries, making it a model for efficiency.
- PA follows with an average delivery time of around 11.44 days. Being an industrial and economic hub in Brazil, Paraná benefits from effective logistics and transportation networks.

**Recommendations:**

- Investigating the specific challenges in this state, such as transportation networks and fulfillment processes, could lead to improvements.
- Fulfillment centers serving in these higher delivery time regions may need to optimize their delivery operations, need to address these challenges to improve customer satisfaction.

15. Find out the top 5 states where the order delivery is 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.

```
with final_delivery as (
select geolocation_state,
extract(date from order_delivered_customer_date) as delivery_date ,
extract(date from order_estimated_delivery_date) as estimated_delivery_date ,
DATE_DIFF(extract(date from order_estimated_delivery_date), extract(date from
order_delivered_customer_date), DAY) as delivery_speed
from `target_casestudy.orders` O
join `target_casestudy.customers` C
on O.customer_id = C.customer_id
join `target_casestudy.geolocation` G
on C.customer_zip_code_prefix = G.geolocation_zip_code_prefix
  )
 select
 geolocation_state,
 round(avg(delivery_speed),2) as avg_delivery_speed
  from
 final_delivery
 group by  geolocation_state
 order by avg_delivery_speed desc
 Limit 5;
```

**Insights:**

- RR, AM, RO, AC, AP are the states where order delivery is fast as compared to the estimated dates.  This suggests that orders are delivered significantly earlier than initially estimated. This efficiency may be attributed to well-optimized logistics and delivery services in the state. This can enhance customer satisfaction.
- Overall, the top 5 states in this list demonstrate effective logistics and delivery operations, leading to orders being consistently delivered well in advance of the estimated delivery dates.

**Recommendation:**

- These insights can be valuable for logistics providers looking to enhance customer satisfaction and operational efficiency in these regions

## 6 Analysis based on the payments:

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

```
 select extract (month from order_purchase_timestamp) as order_month ,
 extract (year from order_purchase_timestamp) as order_year,
 payment_type ,
 count(*) as no_of_orders
from `target_casestudy.orders` O
inner join `target_casestudy.payments`P
on O.order_id = P.order_id
group by 1,2 ,3
order by 2, 1
```

| | | | |
|---|---|---|---|
| 10 | 3 | UPI | 1942 |

**Insights:**

- There are 5 different types of payment methods that have been used, named as credit card, debit card, UPI, voucher and the last one is not defined.
- It's clearly visible that credit cards are the most used payment method across the different months. Post this the second highly used method is UPI then voucher and then Debit card.
- These insights show the variations in the number of orders placed using different payment types for each month

**Recommendation:**

- Can collaborate with different banks to come up with different lucrative credit card offers, marketing campaigns can be optimized with respect to the credit card offers.
- Similarly for UPI, can collaborate with different payment wallets for wallets specific discounts and cashbacks.

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

```
 SELECT
 P.payment_installments,
 COUNT(O.order_id) AS no_of_orders
FROM `target_casestudy.payments` P
INNER JOIN `target_casestudy.orders` O
ON P.order_id = O.order_id
GROUP BY P.payment_installments;
```

**Insights:**

The most preferred payment preference is one shot as per the data. Second one is 2 installments then 3 installments and so on. Order no. decreased as the no. of installments increased only for the 2 exceptions 8th and 10th installments.

**Recommendations:**

To encourage customers to choose certain payment options, consider offering incentives or discounts. For example, you could offer a small discount for customers who select a 3-6-8-12-installment plan to promote the usage of this option.