**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,

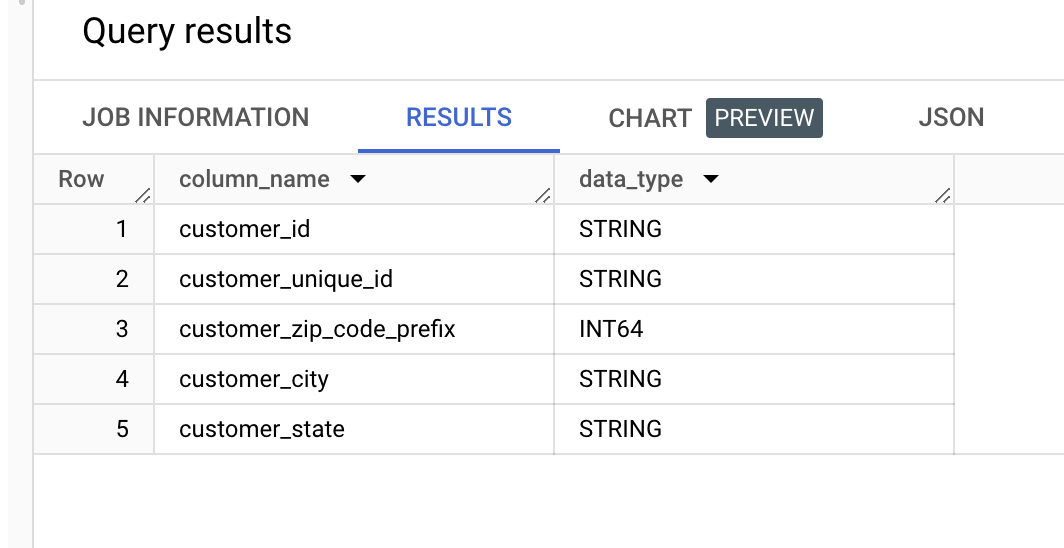
data\_type

FROM

`target-403119.target\_casestudy.INFORMATION\_SCHEMA.COLUMNS`

WHERE

table\_name = 'customers';



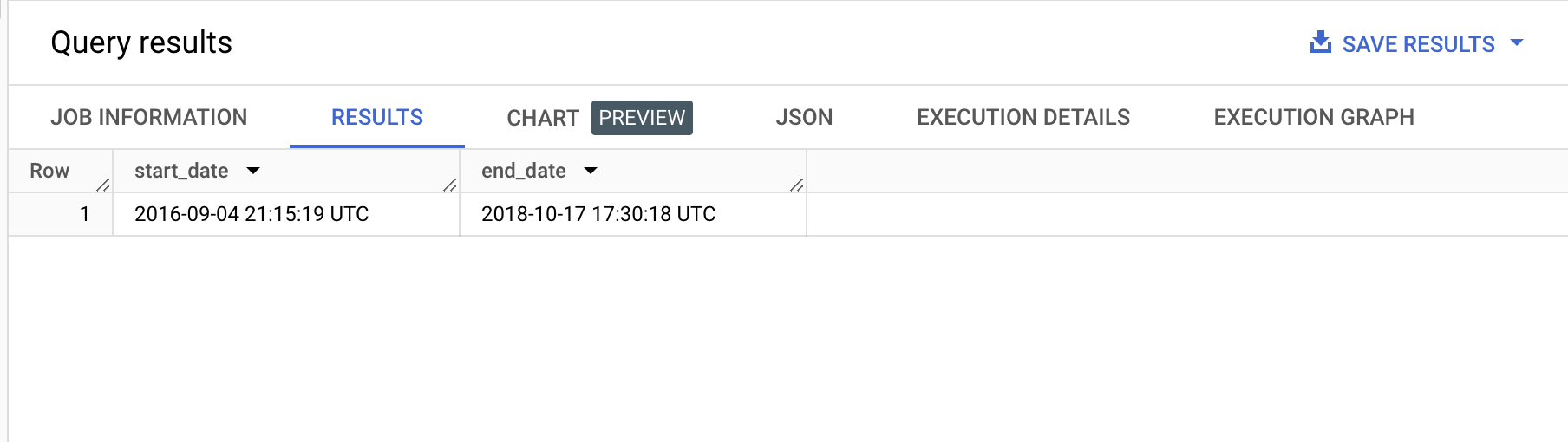
1. Get the time range between which the orders were placed.

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

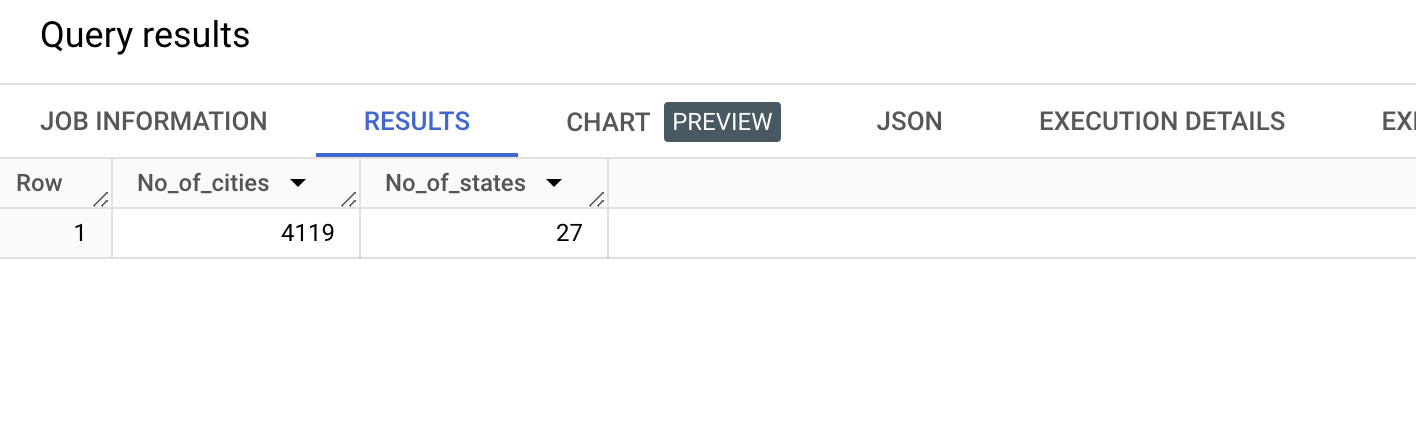
1. Count the Cities & States of customers who ordered during the given period.

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.

**In-depth Exploration:**

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

select

extract(year from order\_purchase\_timestamp) as year ,

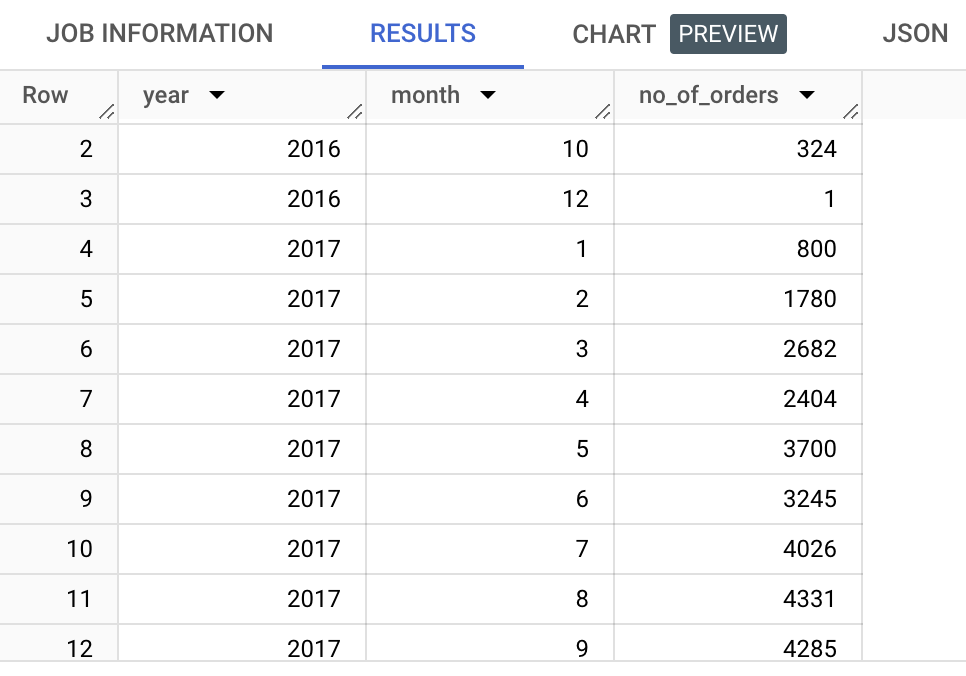
extract(month from order\_purchase\_timestamp) as month,

count(order\_id) as no\_of\_orders

From `target\_casestudy.orders`

group by 1 ,2

order by 1 , 2



**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

1. 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,

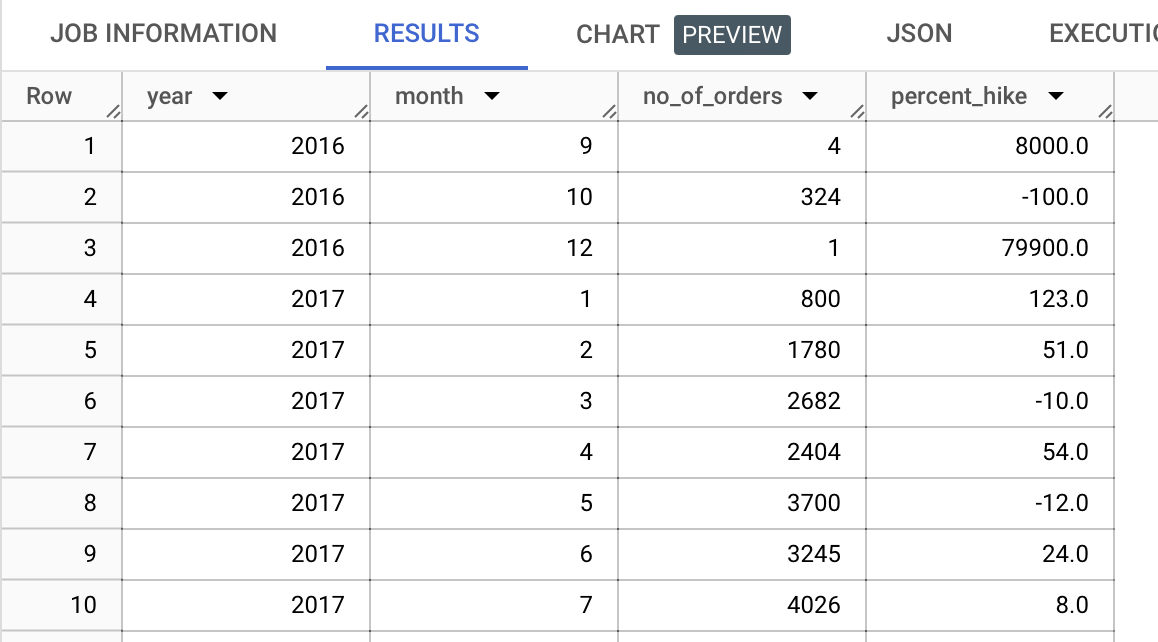
count(order\_id) as no\_of\_orders

from `target\_casestudy.orders`

group by 1 ,2

order by 1 , 2 ) temp

order by year, month



Insights: the 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

1. 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'

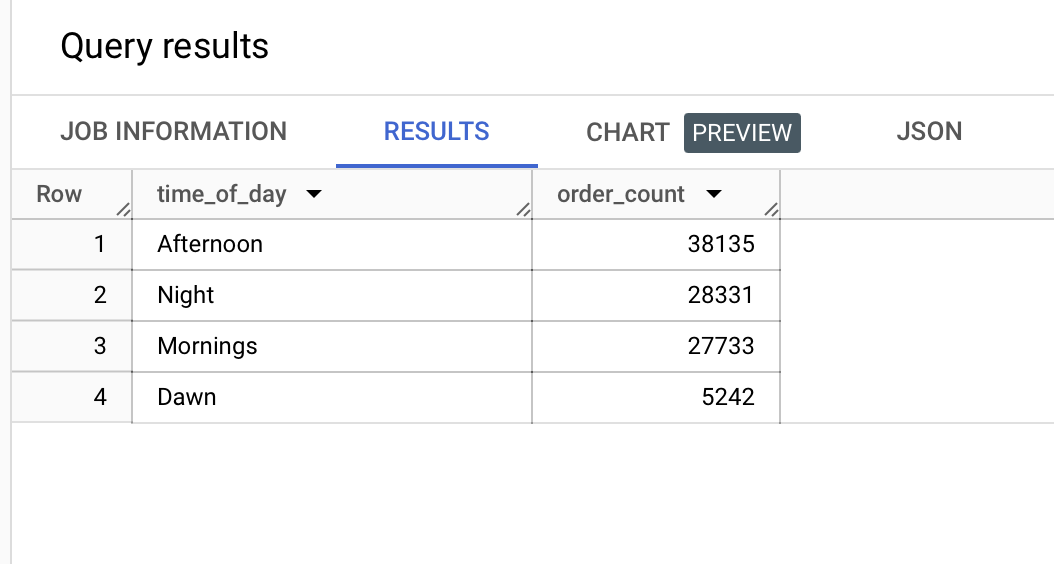
when extract (hour from order\_purchase\_timestamp) between 19 and 23 then 'Night'

end as time\_of\_day

from target\_casestudy.orders ) temp

group by 1

order by 2 desc



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 afternoon time 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.

**Evolution of E-commerce orders in the Brazil region:**

1. 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

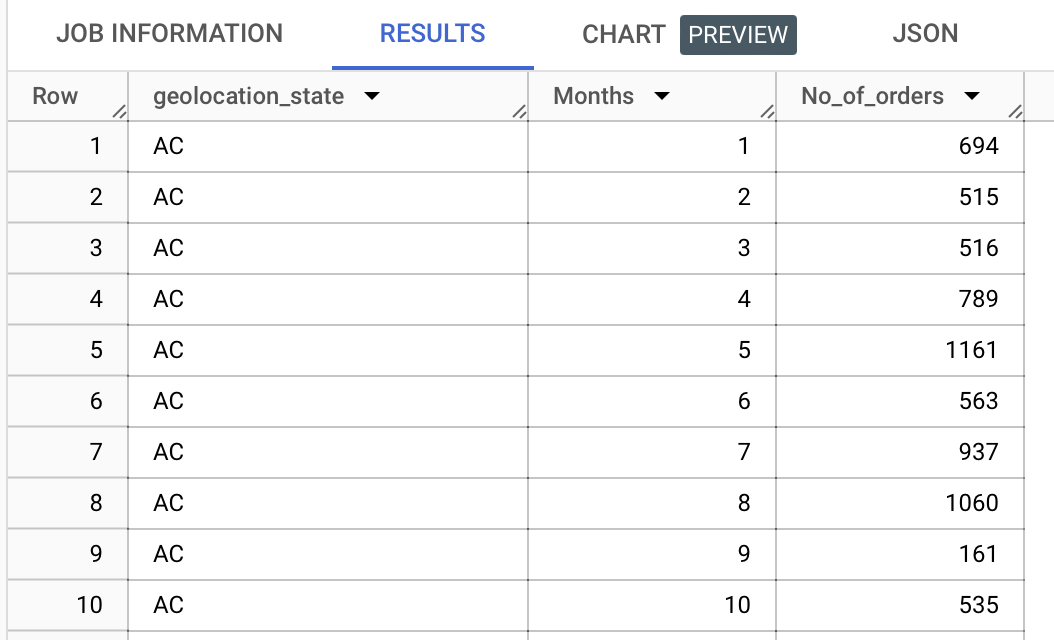
FROM `target\_casestudy.orders` O

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;



SELECT geolocation\_state ,

COUNT (\*) as No\_of\_orders

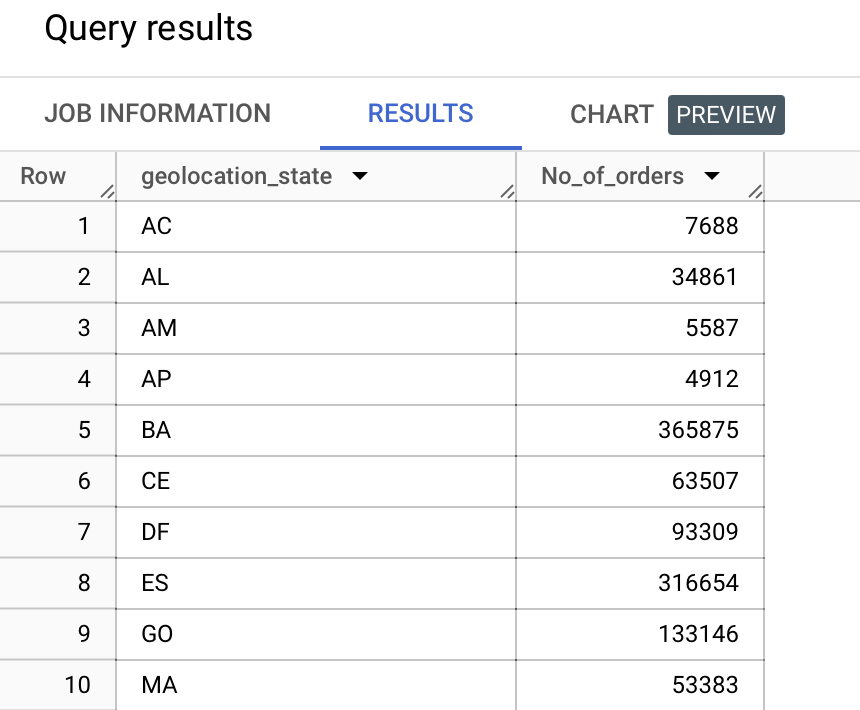
FROM `target\_casestudy.orders` O

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



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.

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

SELECT

geolocation\_state,

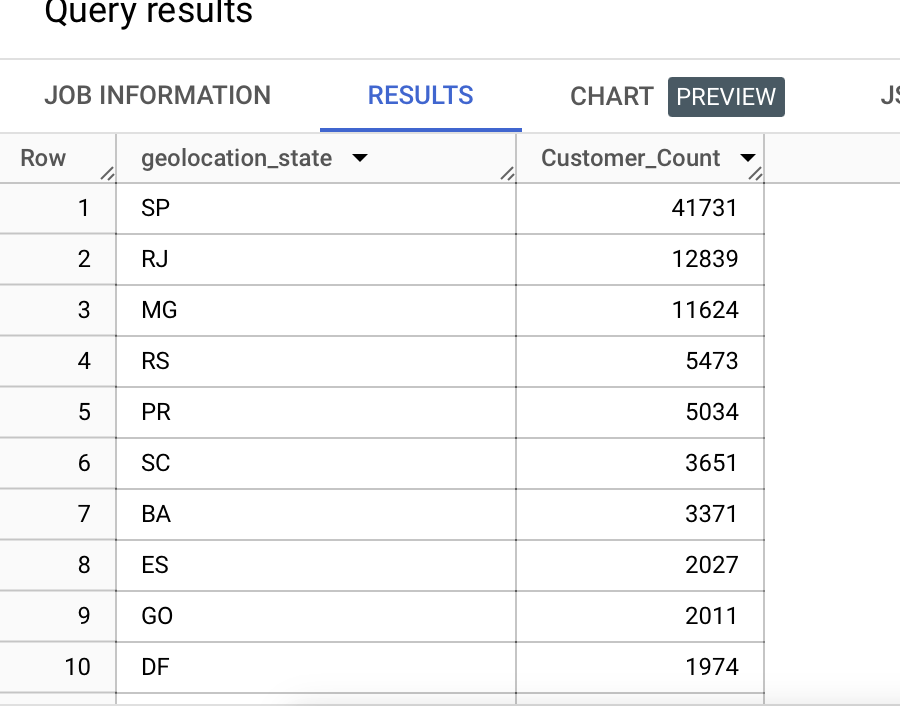
COUNT(DISTINCT customer\_id) as Customer\_Count

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.

**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 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

)

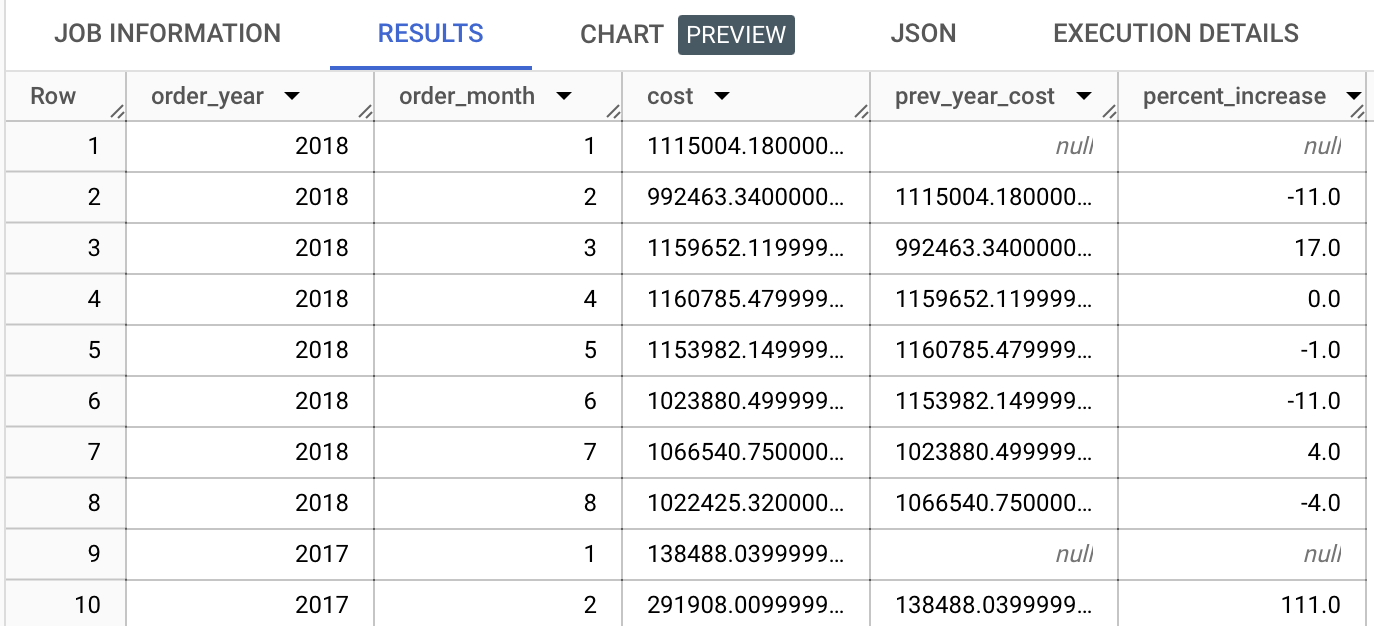
select order\_year , order\_month , cost ,

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



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,

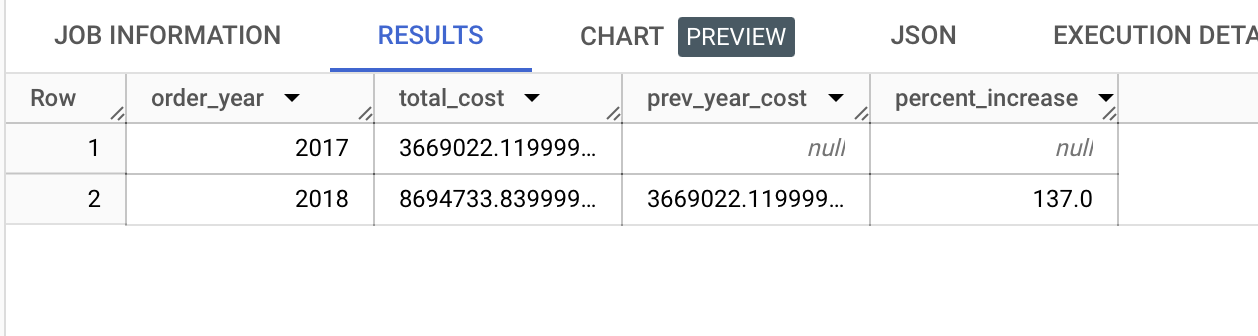
LAG(SUM(cost), 1) OVER (ORDER BY order\_year) AS prev\_year\_cost,

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

1. 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

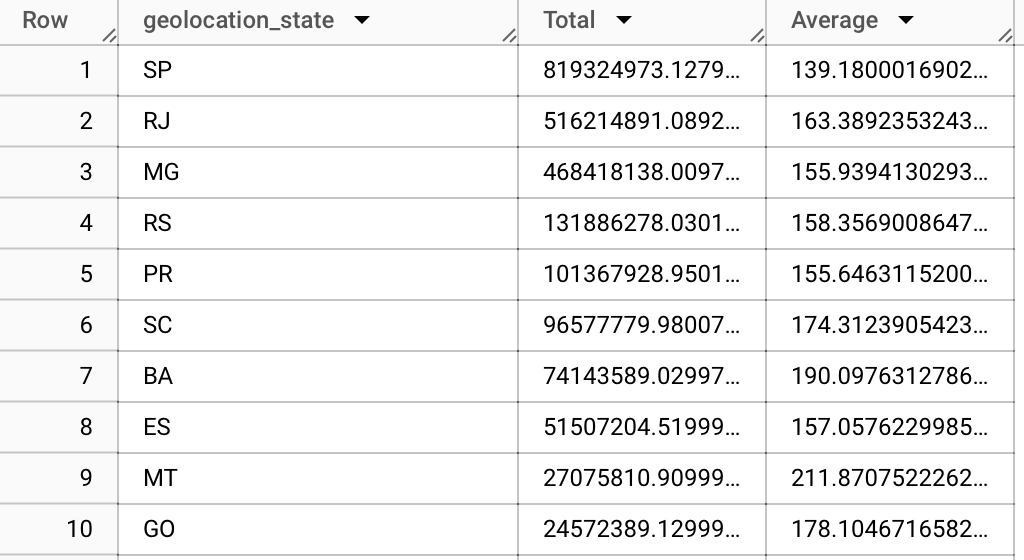
on O.customer\_id= C.customer\_id

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



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.

1. 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

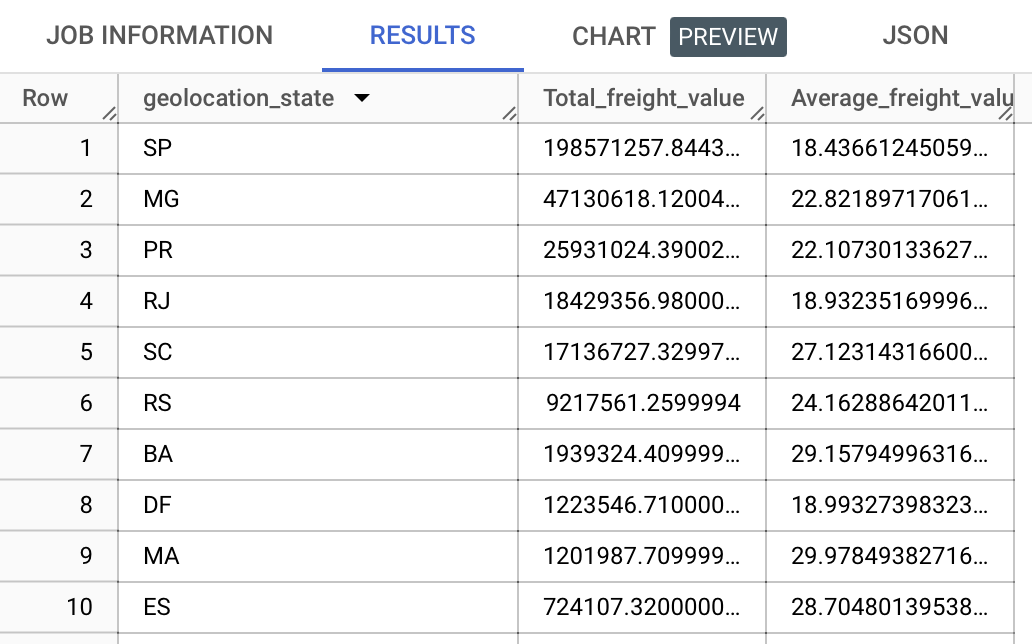
on OI.seller\_id = S.seller\_id

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.

**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.  
   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` )

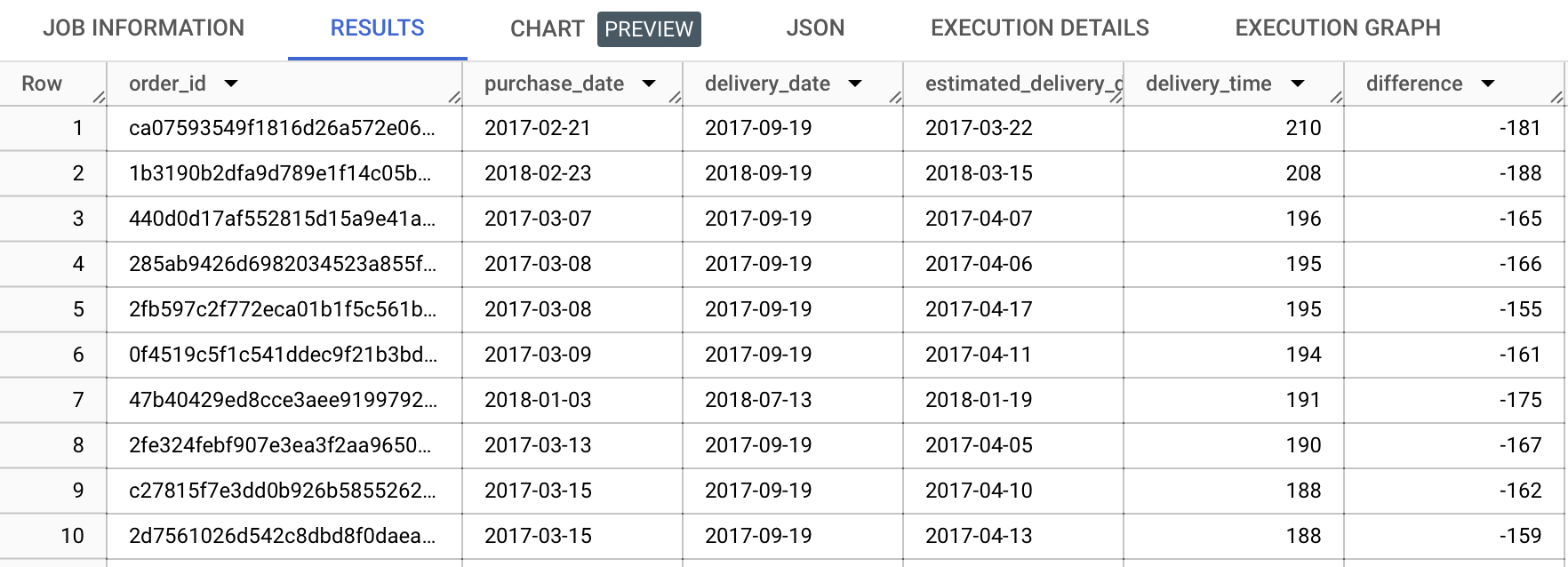
select \* ,

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



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.

1. 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

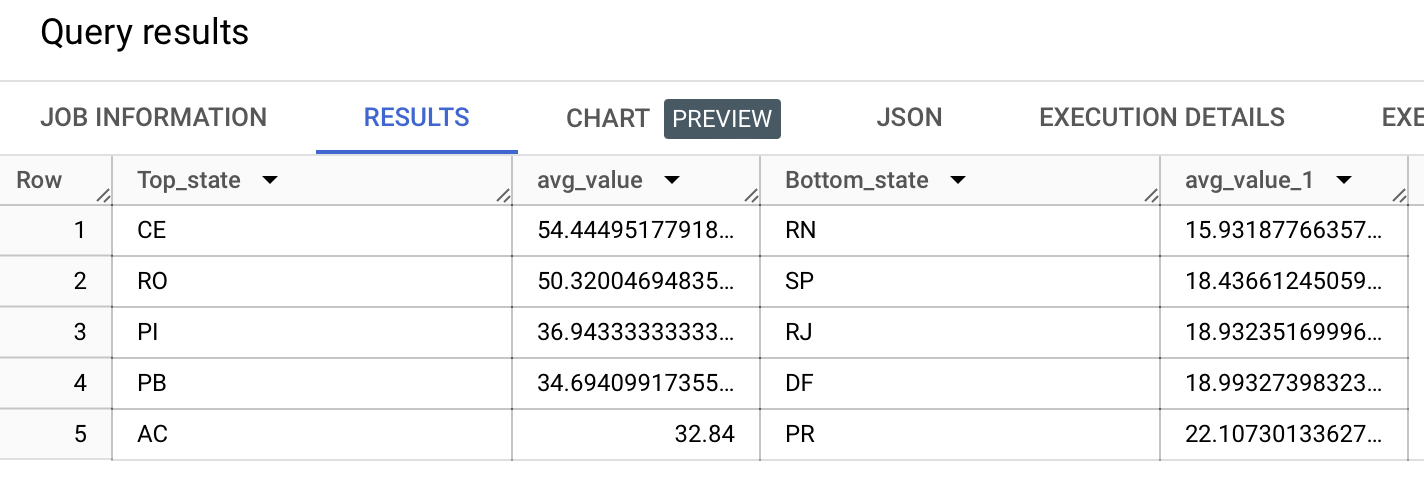
(select geolocation\_state , avg\_value , row\_number()over(order by avg\_value asc) R2

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



1. 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

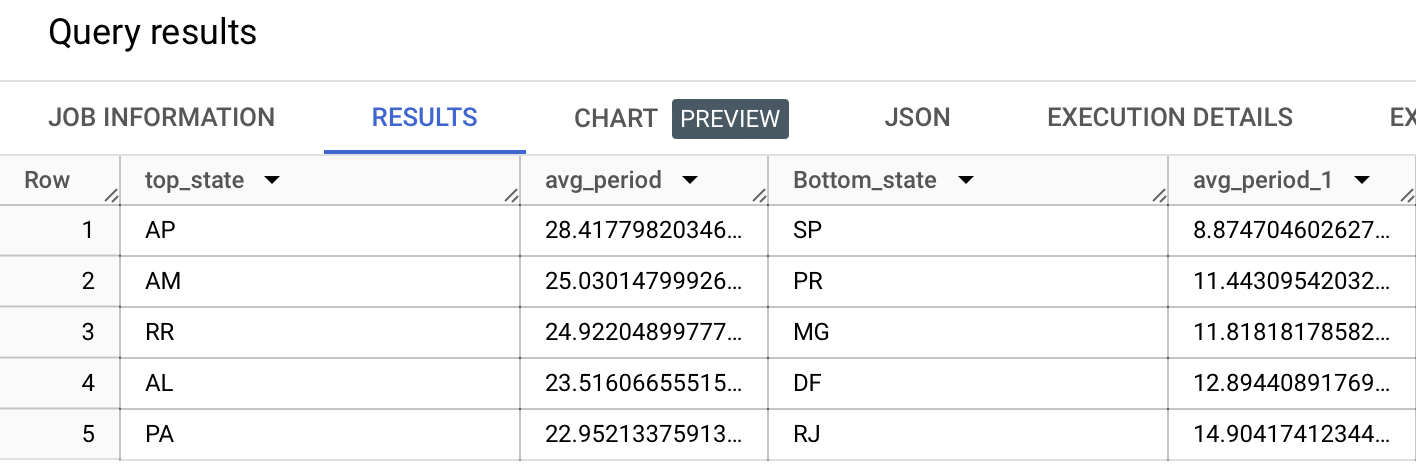
from delivery\_data

group by 1

order by 2 asc

limit 5 ) B

on A.R1= B.R2



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.

1. 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

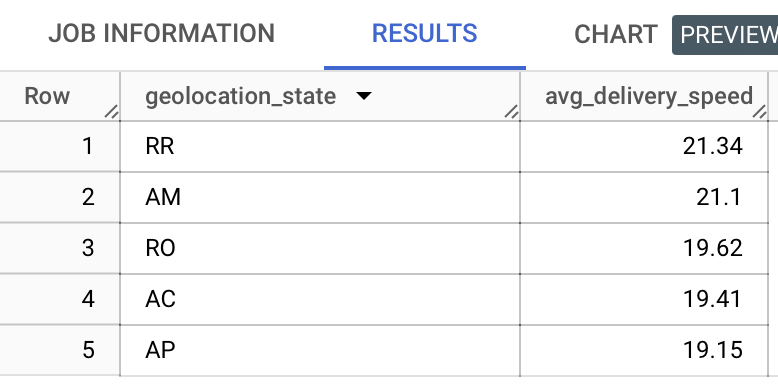
group by

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

**Analysis based on the payments:**

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