1. **Import the dataset and do usual exploratory analysis steps like checking the structure & characteristics of the dataset**

A) Data type of columns in a table

Create a new project →Create dataset→Create Table →Then click on schema

B) Time period for which the data is given

Start from (2016-09-04 21:15:19 UTC)

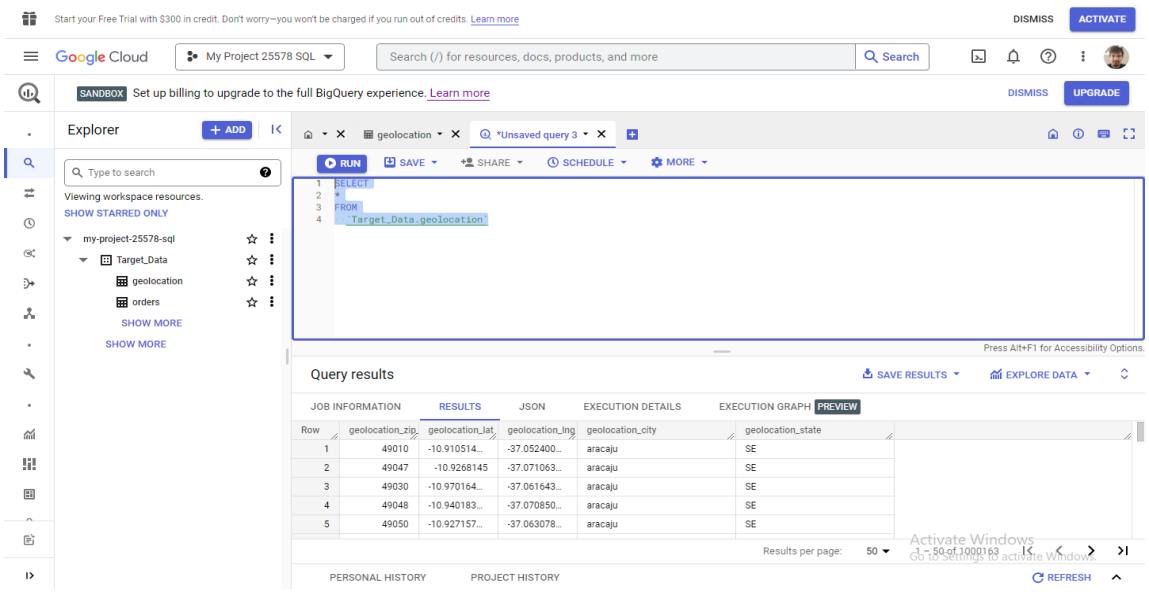
```
SELECT
   MIN(order_purchase_timestamp)
FROM ‘Target_Data.orders`
```

Latest is (2018-10-17 17:30:18 UTC)

```
SELECT
   MAX(order_purchase_timestamp)
FROM `Target_Data.orders`
```

C) Cities and States of customers ordered during the given period
For that we need to create a table of geolocation

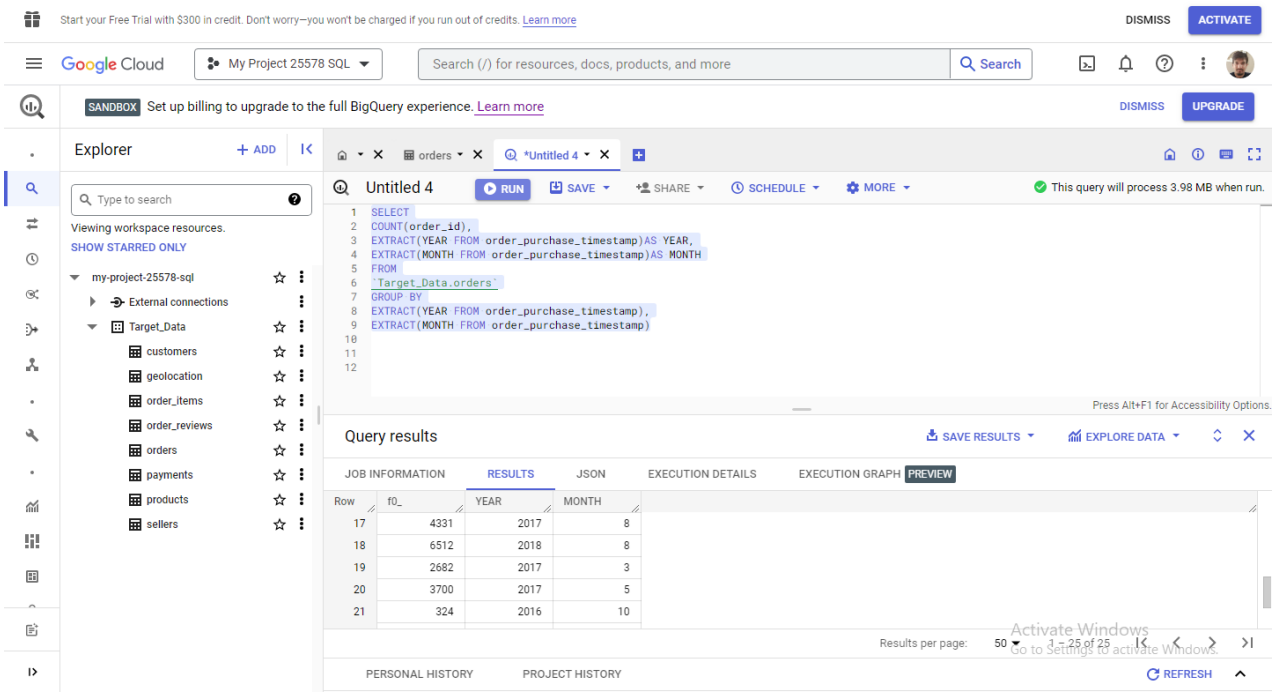
```
SELECT
       *
FROM `Target_Data.geolocation`
```



2. In-depth Exploration:

A) 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?

```
SELECT
COUNT(order_id),
EXTRACT(YEAR FROM order_purchase_timestamp)AS YEAR,
EXTRACT(MONTH FROM order_purchase_timestamp)AS MONTH
FROM
`Target_Data.orders`
GROUP BY
EXTRACT(YEAR FROM order_purchase_timestamp),
EXTRACT(MONTH FROM order_purchase_timestamp)
```

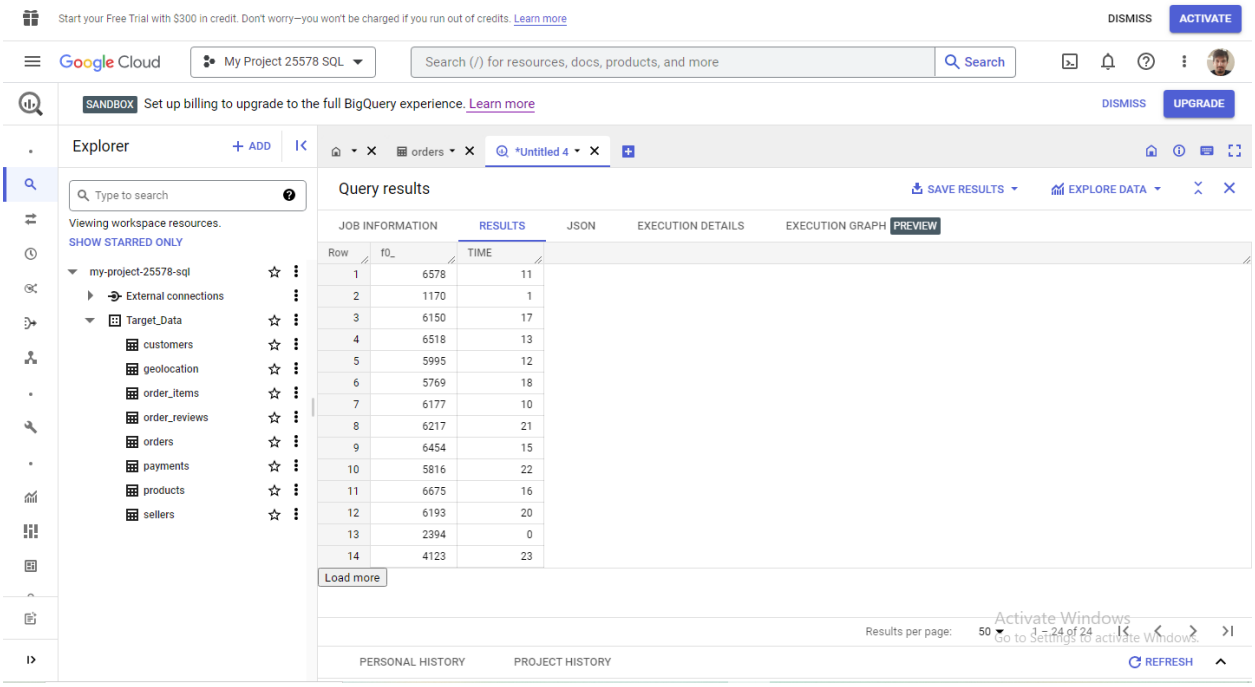


Yes there is a clear growing trend (YOY) as the numbers of order are increasing.

And also seasonality peaks can be observed in the months of November,December and January as these are festive and holiday months .

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

```
SELECT

COUNT(order_id),
EXTRACT(HOUR FROM order_purchase_timestamp) AS TIME
FROM
`Target_Data.orders`
GROUP BY
EXTRACT(HOUR FROM order_purchase_timestamp)
```

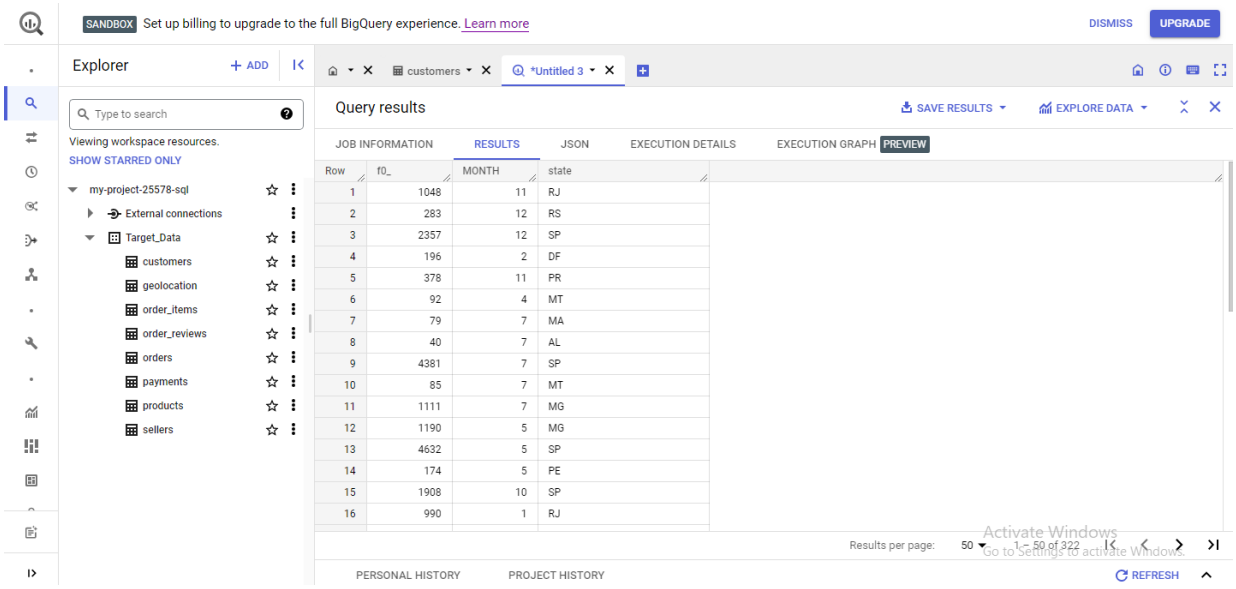


Most of the orders are placed in afternoon and night time as people of brazil must be having free time at these time intervals.

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

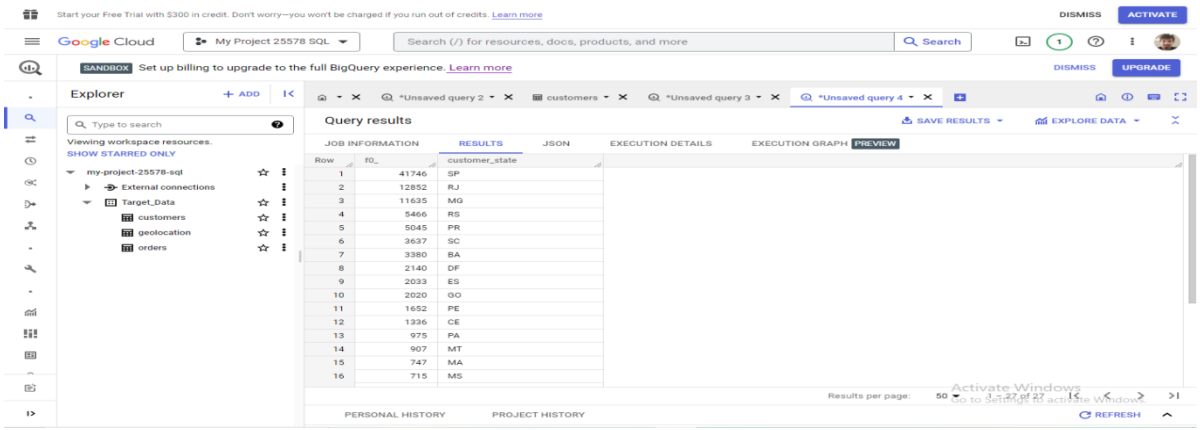
A) Get month on month orders by states

```
SELECT count(order_id),EXTRACT (MONTH FROM order_purchase_timestamp) AS MONTH,customer_state AS state

FROM`Target_Data.orders` AS o left join `Target_Data.customers` AS c ON o.customer_id=c.customer_id
GROUP BY EXTRACT (MONTH FROM order_purchase_timestamp),customer_state
```



B) Distribution of customers across the states in Brazil

```
SELECT
    COUNT(*),
    customer_state
FROM`Target_Data.customers`
GROUP BY
    customer_state
ORDER BY
    COUNT(*) DESC
```



It seems that most of the customers are located in the richest state of Brazil. As it is obvious that the purchasing power is high of the people of São Paulo and more number of customers is because it has the largest population.

Note that Roraima with lowest GDP contribution is at the bottom when it comes to number of customers.

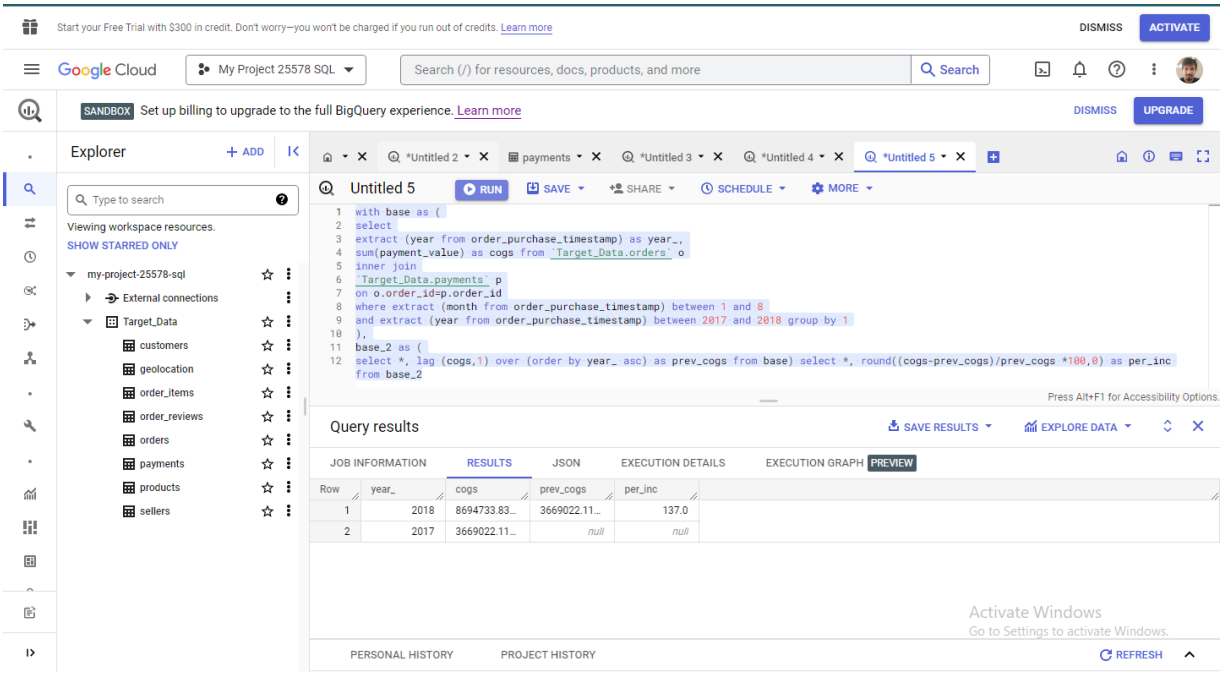
So it can be concluded that the GDP contribution of a state has direct relation with the number of customers.

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

A) 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.

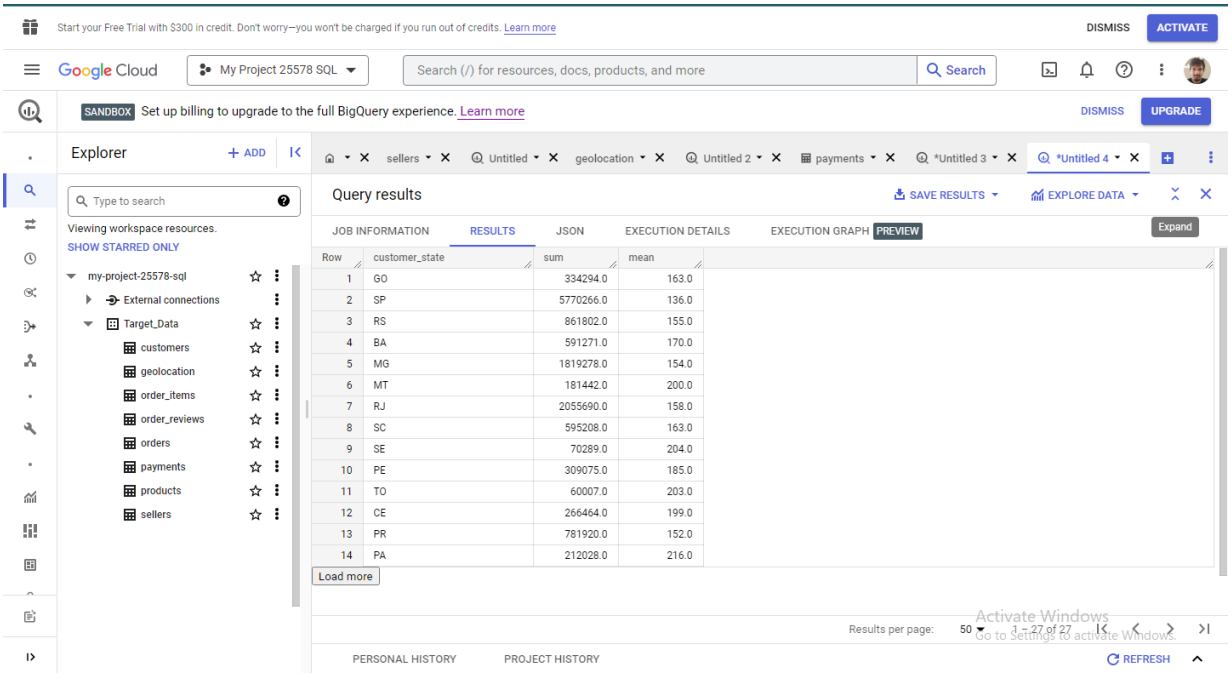
```
with base as (
select
extract (year from order_purchase_timestamp) as year_,
sum(payment_value) as cogs from `Target_Data.orders` o
inner join
`Target_Data.payments` p
on o.order_id=p.order_id
where extract (month from order_purchase_timestamp) between 1 and 8
and extract (year from order_purchase_timestamp) between 2017 and 2018 group by 1
),
```

```
base_2 as (
select *, lag (cogs,1) over (order by year_ asc) as prev_cogs from base) select *, round((c
ogs-prev_cogs)/prev_cogs *100,0) as per_inc from base_2
```



B) Mean & Sum of price and freight value by customer state

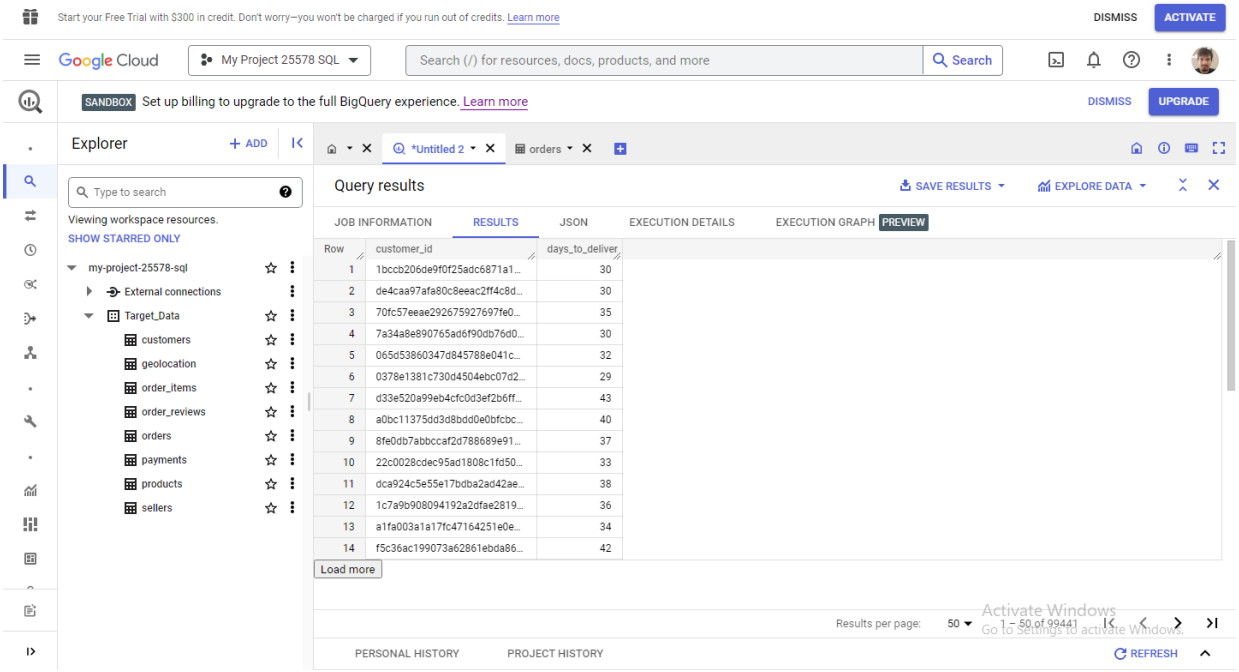
```
select c.customer_state ,round(sum(payment_value)) as sum,round(avg(payment_value)) as mean
from `Target_Data.customers` c join `Target_Data.orders` o on c.customer_id = o.customer_id join `Target_
Data.payments` p on o.order_id = p.order_id
where o.order_status = 'delivered'
group by 1
```



5. Analysis on sales, freight and delivery time

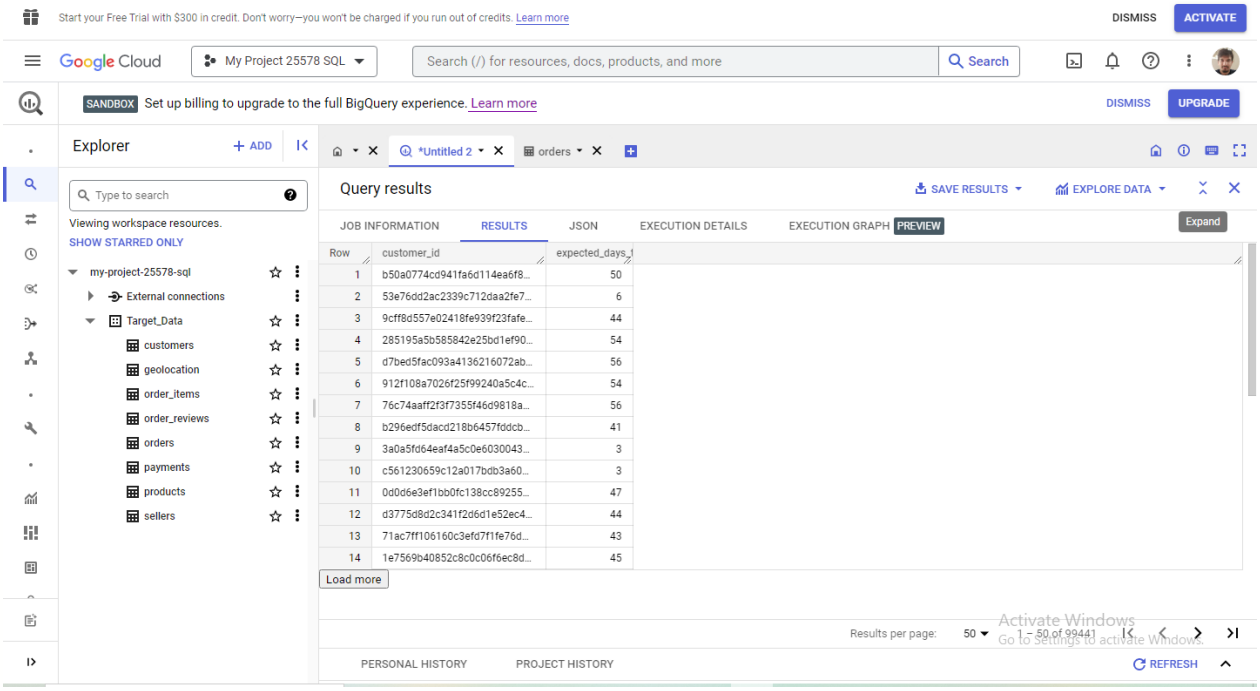
A) Calculate days between purchasing, delivering and estimated delivery

```
SELECT
customer_id,
DATE_DIFF(order_delivered_customer_date, order_purchase_timestamp, day) AS days_to_deliver
FROM `Target_Data.orders`
```



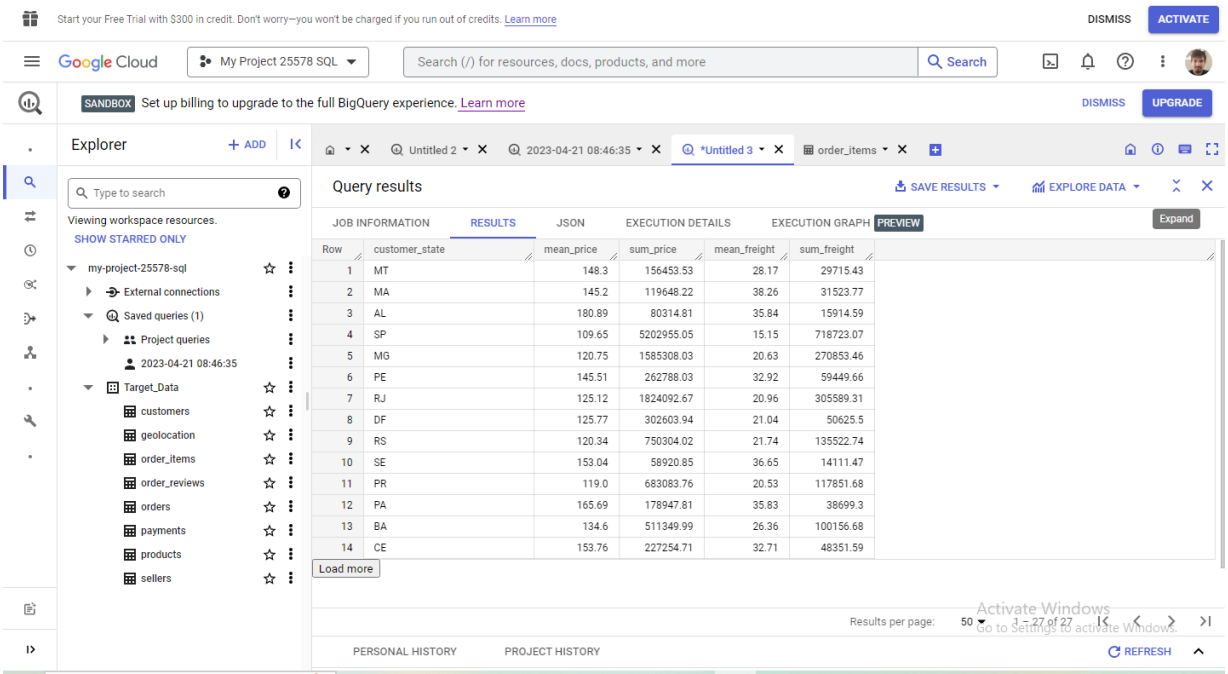
B) Find time_to_delivery & diff_estimated_delivery. Formula for the same given below:

```
SELECT
customer_id,
DATE_DIFF(order_estimated_delivery_date, order_purchase_timestamp, day) AS expected_days_to
_deliver
FROM `Target_Data.orders`
```



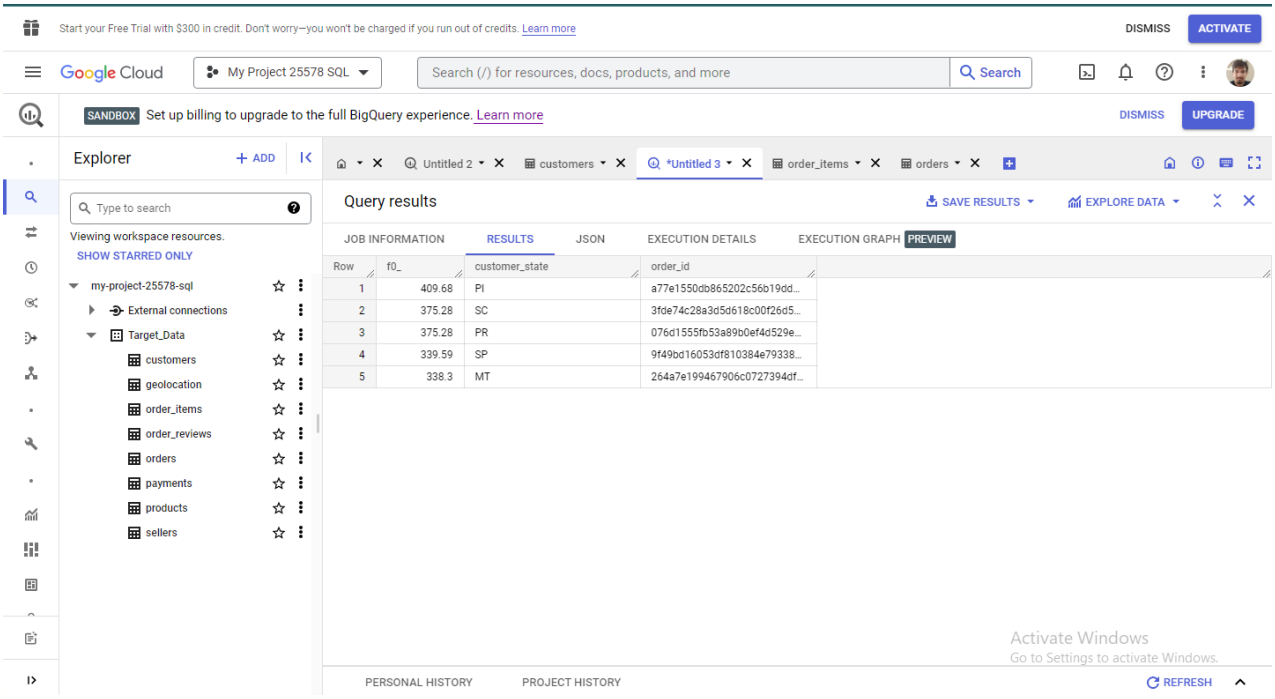
C) Group data by state, take mean of freight_value, time_to_delivery, diff_estimated_delivery

```
SELECT c.customer_state,

ROUND(AVG(oi.price), 2) AS mean_price,

ROUND(SUM(oi.price), 2) AS sum_price,

ROUND(AVG(oi.freight_value), 2) AS mean_freight,

ROUND(SUM(oi.freight_value), 2) AS sum_freight

FROM `Target_Data.customers` AS c

JOIN `Target_Data.orders` AS o

ON c.customer_id = o.customer_id

JOIN `Target_Data.order_items` AS oi

ON o.order_id = oi.order_id

GROUP BY c.customer_state
```

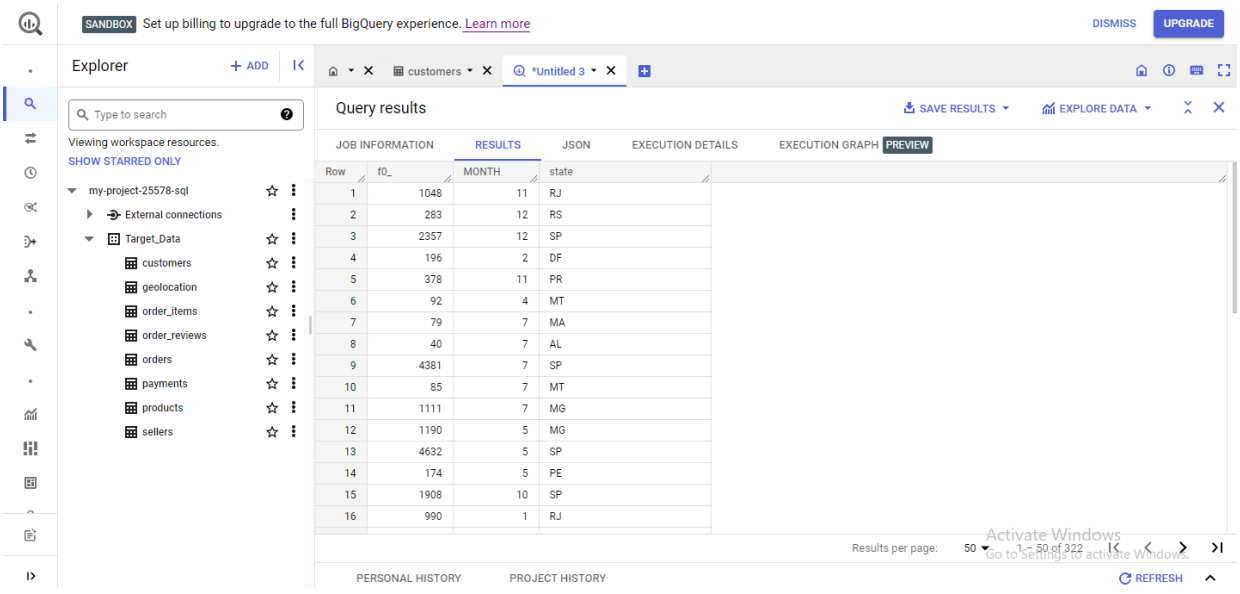


D) Top 5 states with highest/lowest average freight value - sort in desc/asc limit 5

```
SELECT
AVG(freight_value),
c.customer_state,
o.order_id
FROM
`Target_Data.orders`AS o join `Target_Data.order_items` AS oi
ON o.order_id=oi.order_id join `Target_Data.customers` AS c
ON o.customer_id = c.customer_id
GROUP BY
customer_state,
order_id
ORDER BY
AVG(freight_value) DESC
LIMIT 5
```

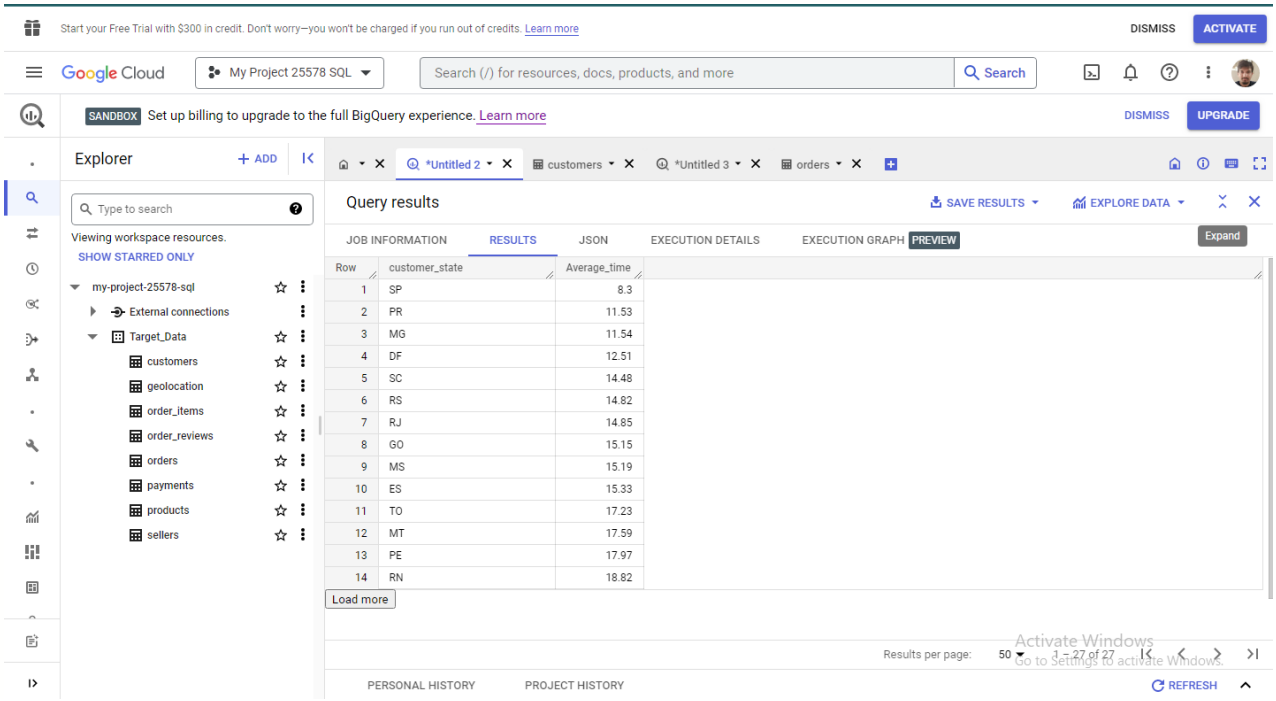



```sql
SELECT
AVG(freight_value),
c.customer_state,
o.order_id
FROM
`Target_Data.orders`AS o join `Target_Data.order_items` AS oi
ON o.order_id=oi.order_id join `Target_Data.customers` AS c
ON o.customer_id = c.customer_id
GROUP BY
customer_state,
order_id
ORDER BY
AVG(freight_value)
LIMIT 5
```



E) Top 5 states with highest/lowest average time to delivery

```
SELECT
customer_state,
ROUND(AVG(DATE_DIFF(order_delivered_customer_date, order_purchase_timestamp, day)),2) Avera
ge_time
FROM
`Target_Data.orders`AS o join `Target_Data.customers` AS c
ON  o.customer_id = c.customer_id
GROUP BY
customer_state
ORDER BY
AVG(DATE_DIFF(order_delivered_customer_date, order_purchase_timestamp, day))
```



F) Top 5 states where delivery is really fast/ not so fast compared to estimated date

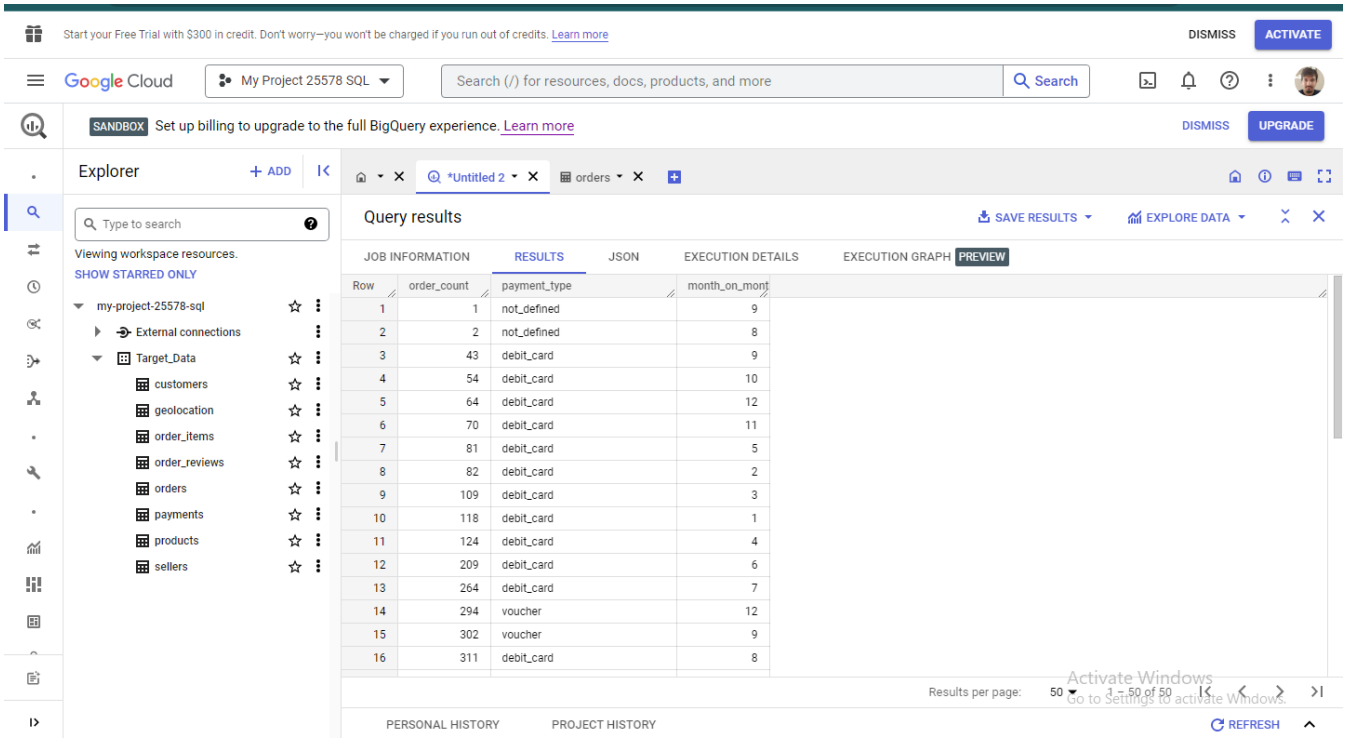
```
WITH state_avg_estimated_delivery AS (

SELECT

c.customer_state,

ROUND(AVG(DATE_DIFF(o.order_estimated_delivery_date, o.order_delivered_

customer_date,DAY)), 2) AS mean_diff_estimated_delivery

FROM `Target_Data.orders` AS o

JOIN `Target_Data.customers` AS c

ON o.customer_id = c.customer_id

GROUP BY c.customer_state

HAVING COUNT(*) > 1000 -

- only considering states with more than 1000 orders

)
```

```
(SELECT
"Top 5 States where delivery is really fast" AS title,
customer_state,
mean_diff_estimated_delivery
FROM state_avg_estimated_delivery
ORDER BY mean_diff_estimated_delivery ASC
LIMIT 5)
UNION ALL
(SELECT
"Top 5 States where delivery is not so fast" AS title,
customer_state,
mean_diff_estimated_delivery
FROM state_avg_estimated_delivery
ORDER BY mean_diff_estimated_delivery DESC
LIMIT 5)
```

6.Payment type analysis:

A) Month over Month count of orders for different payment types

```
SELECT COUNT(o.order_id) AS order_count, p.payment_type,
EXTRACT (MONTH FROM order_purchase_timestamp) AS month_on_month
 FROM`Target_Data.orders` AS o join `Target_Data.payments` AS p ON o.order_id = p.order_id
GROUP BY payment_type, month_on_month
ORDER BY COUNT(o.order_id)
```



B) Count of orders based on the no. of payment installments

```
SELECT
COUNT(o.order_id) AS no_of_orders,
p.payment_installments
FROM `Target_Data.orders` AS o join `Target_Data.payments` AS p
ON o.order_id = p.order_id
GROUP BY
p.payment_installments
ORDER BY
COUNT(o.order_id)
```