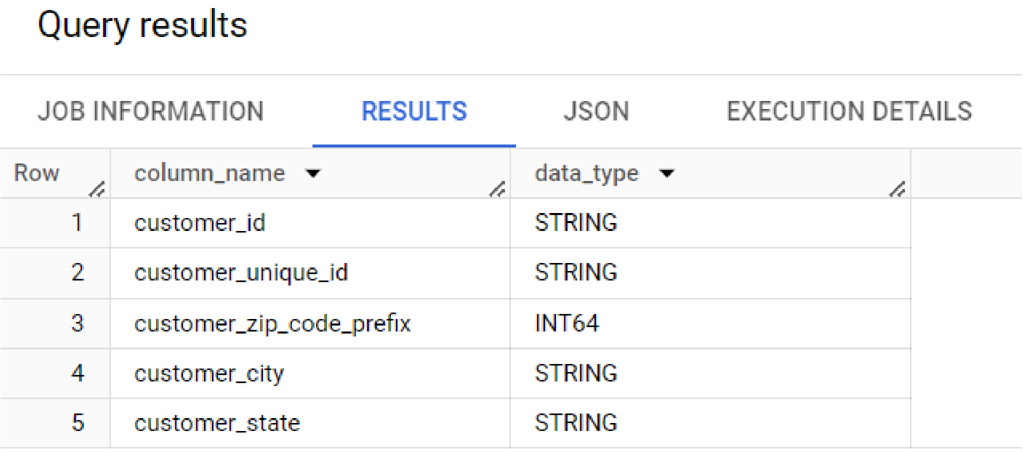
Target Case Study

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.

Solution: -

```
SELECT column_name, data_type
FROM `boxwood-magnet-396716.CaseStudyTarget.INFORMATION_SCHEMA.COLUMNS`
WHERE table_name = 'customers';
```

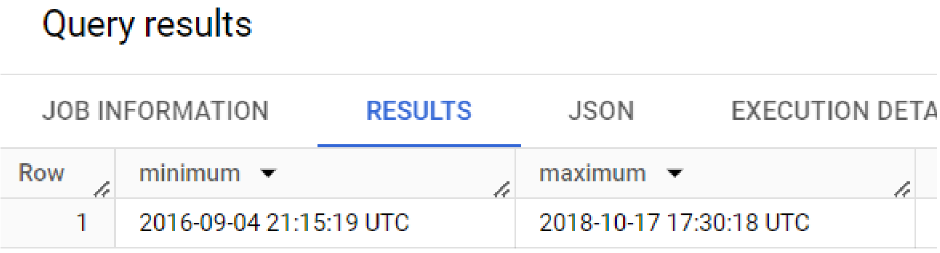


Row	column_name	data_type
1	customer_id	STRING
2	customer_unique_id	STRING
3	customer_zip_code_prefix	INT64
4	customer_city	STRING
5	customer_state	STRING

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

Solution: -

```
select min(order_purchase_timestamp) as minimum,
max(order_purchase_timestamp) as maximum
from `CaseStudyTarget.orders`;
```



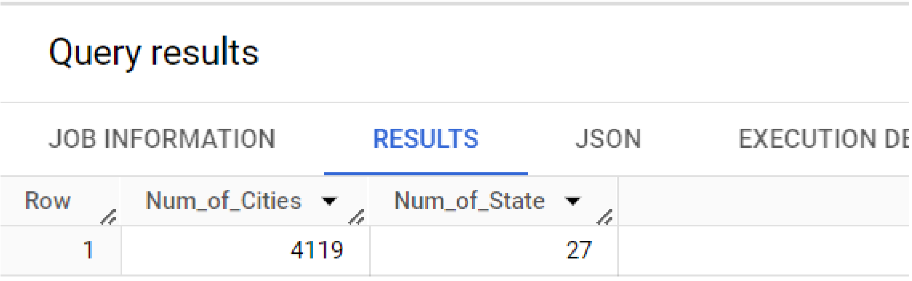
Row	minimum	maximum
1	2016-09-04 21:15:19 UTC	2018-10-17 17:30:18 UTC

- The sale started on 4th September 2016 and the cycle ended at 17th October 2018.

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

Solution: -

```
select count(distinct c.customer_city) as Num_of_Cities ,count(distinct c.customer_state)
as Num_of_State
from `CaseStudyTarget.customers` c join `CaseStudyTarget.orders` o
on c.customer_id=o.customer_id;
```



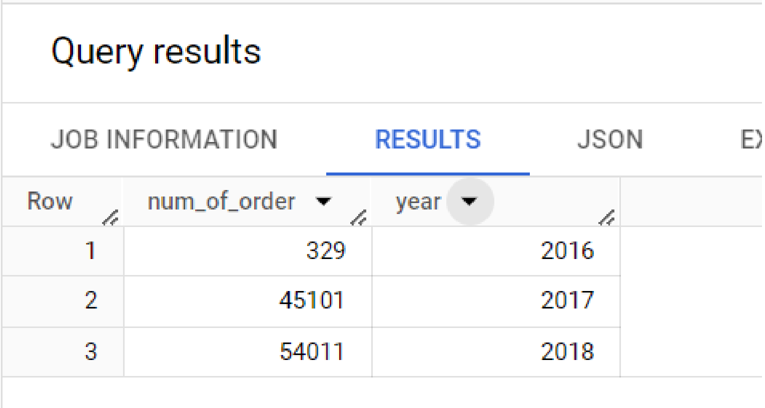
Row	Num_of_Cities	Num_of_State
1	4119	27

- The orders were placed from 4119 different cities and 27 different states

2. In-depth Exploration

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

```
select count(order_id) as num_of_order,
extract(year from order_purchase_timestamp) as year
from `CaseStudyTarget.orders`
group by 2
order by 2;
```

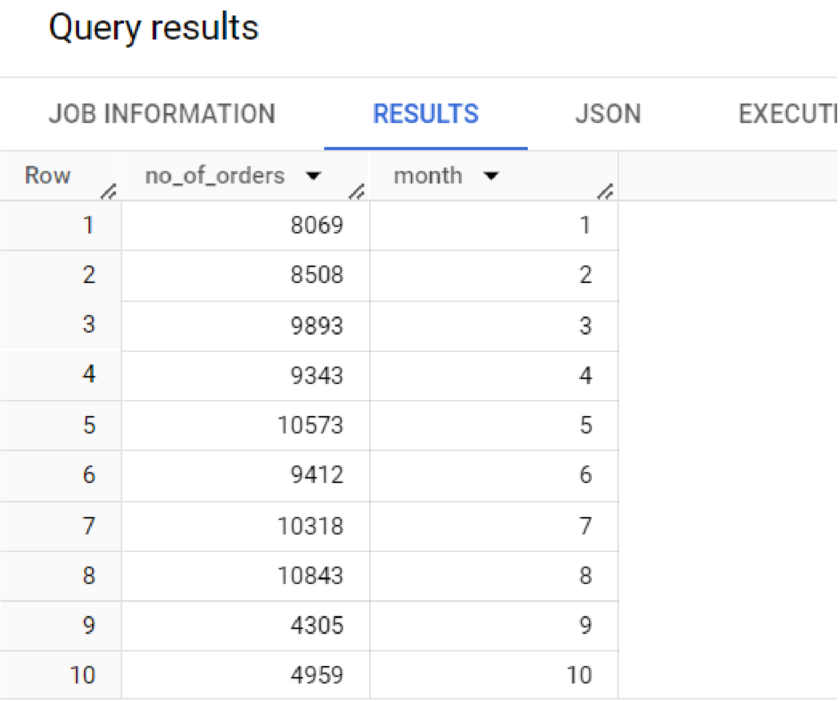


Row	num_of_order	year
1	329	2016
2	45101	2017
3	54011	2018

- A noticeable trend of increasing order numbers is observed over the years

2. Can we see some kind of monthly seasonality in terms of the no. of orders being placed?
Solution: -

```
with sidcte as(
select count(order_id) as no_of_orders ,
extract(month from order_purchase_timestamp) as month from `CaseStudyTarget.orders`
group by extract(month from order_purchase_timestamp)
)
select * from sidcte
order by month;
```


Query results

Row	no_of_orders	month
1	8069	1
2	8508	2
3	9893	3
4	9343	4
5	10573	5
6	9412	6
7	10318	7
8	10843	8
9	4305	9
10	4959	10

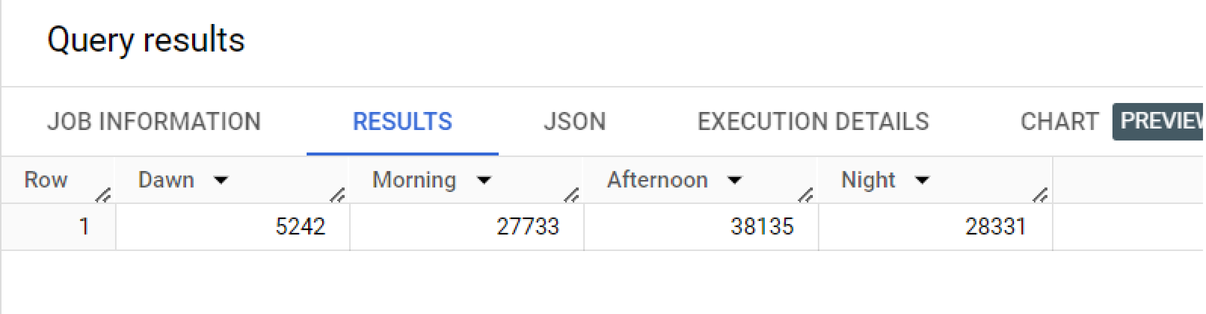
- ➢ It has been observed that the number of orders has been suddenly decreased after August.

3. 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

Solution: -

```
select
sum(
case
when extract(hour from order_purchase_timestamp) between 0 and 6 then 1 else 0
end
) as Dawn,
sum(
case
when extract(hour from order_purchase_timestamp) between 7 and 12 then 1 else 0
end
) as Morning,
sum(
case
when extract(hour from order_purchase_timestamp) between 13 and 18 then 1 else 0
end
) as Afternoon,
sum(
case
when extract(hour from order_purchase_timestamp) between 19 and 23 then 1 else 0
end
) as Night
from `CaseStudyTarget.orders`
```



Row	Dawn	Morning	Afternoon	Night
1	5242	27733	38135	28331

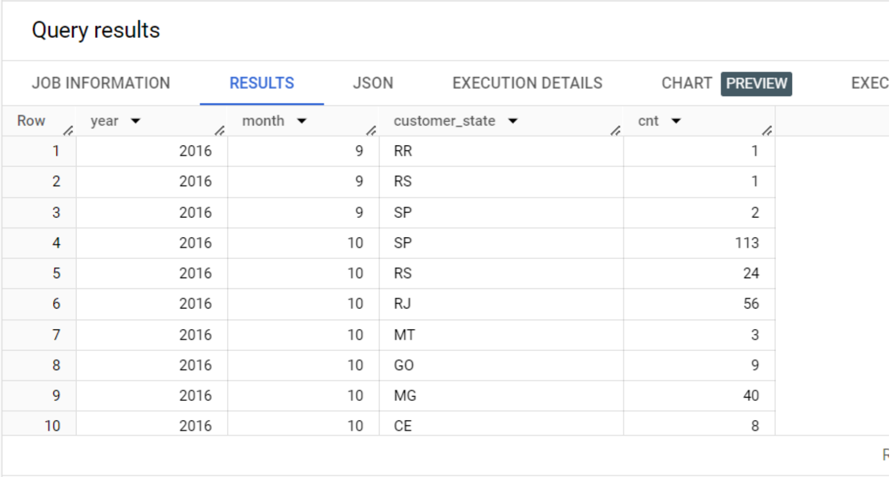
- ➢ Brazilian Customers mostly placed their orders in the afternoon i.e. from 1pm till 6:59 pm

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

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

Solution: -

```
select extract(year  from order_purchase_timestamp) as year,
extract(month  from order_purchase_timestamp) as month,
customer_state,
count(order_id) as cnt
from CaseStudyTarget.orders  inner join CaseStudyTarget.customers  using(customer_id)
group by 1,2,3
order by year,month;
```

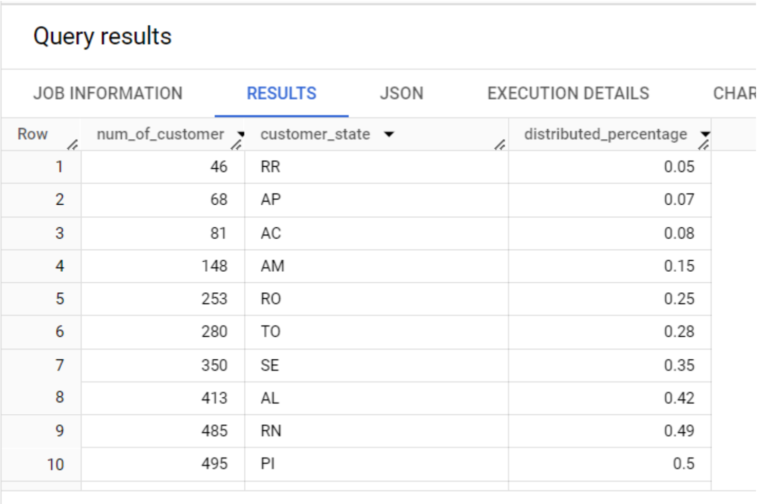


Row	year	month	customer_state	cnt
1	2016	9	RR	1
2	2016	9	RS	1
3	2016	9	SP	2
4	2016	10	SP	113
5	2016	10	RS	24
6	2016	10	RJ	56
7	2016	10	MT	3
8	2016	10	GO	9
9	2016	10	MG	40
10	2016	10	CE	8

- ➢ It has been observed that the greatest number of orders for each month has been placed from the State ‘SP’

2. How are the customers distributed across all the states?
Solution: -

```
with sidcte as(
select count(customer_id) as num_of_customer,customer_state from
`CaseStudyTarget.customers`
group by customer_state
)
select sidcte.num_of_customer,customer_state,
round(num_of_customer/(select count(customer_id) from `CaseStudyTarget.customers`)*100,2)
as distributed_percentage
from sidcte
order by sidcte.num_of_customer;
```



Row	num_of_customer	customer_state	distributed_percentage
1	46	RR	0.05
2	68	AP	0.07
3	81	AC	0.08
4	148	AM	0.15
5	253	RO	0.25
6	280	TO	0.28
7	350	SE	0.35
8	413	AL	0.42
9	485	RN	0.49
10	495	PI	0.5

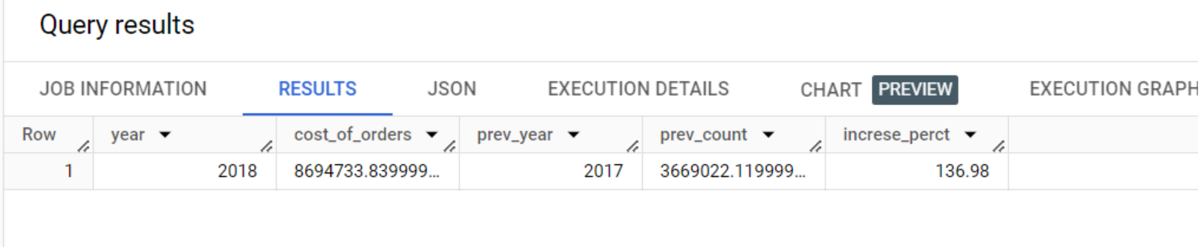
- It has been observed that the customer is not equally distributed major chunk of customer are from the state ‘SP’ which isn’t a good sign for the sellers in other state

4. **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 year 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.
Solution: -

```
with sidcte as(
select sum(p.payment_value) as cost_of_orders,extract(year from order_purchase_timestamp)
as year
from `CaseStudyTarget.orders`o join `CaseStudyTarget.payments` p
on o.order_id=p.order_id
where extract(month from order_purchase_timestamp) between 1 and 8
group by extract(year from order_purchase_timestamp)
)
select *,round(((cost_of_orders-prev_count)/prev_count)*100,2) as increse_perct from(
select year,cost_of_orders,
lag(year) over(order by sidcte.year) as prev_year,
lead(cost_of_orders) over(order by cost_of_orders desc) as prev_count
from sidcte) t
```

```
where t.prev_count is not null
order by cost_of_orders desc;
```

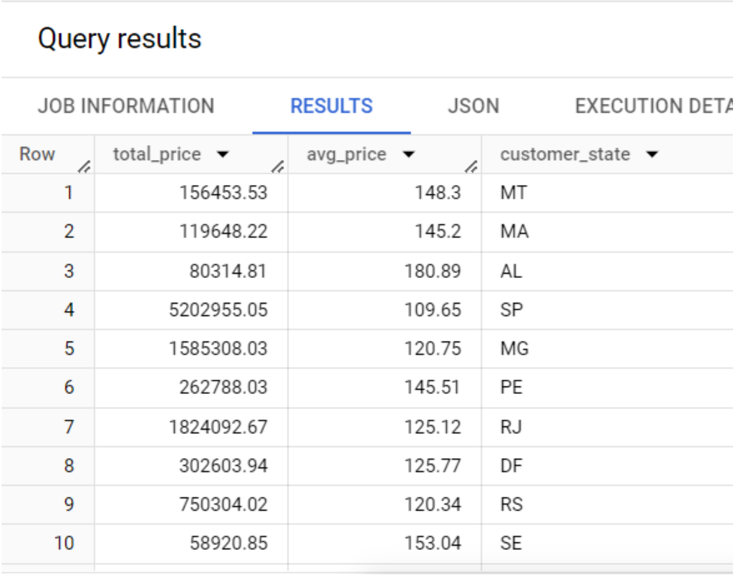


Row	year	cost_of_orders	prev_year	prev_count	increse_perct
1	2018	8694733.839999...	2017	3669022.119999...	136.98

- ➢ The cost of orders experienced a increase of nearly 137% from January to August in the year 2018 compared to 2017.

2.Calculate the Total & Average value of order price for each state.
Solution: -

```
select round(sum(ot.price),2) as total_price,
round(avg(ot.price),2) as avg_price,
c.customer_state
from `CaseStudyTarget.customers` c
join
`CaseStudyTarget.orders` o on c.customer_id=o.customer_id
join
`CaseStudyTarget.order_items` ot on ot.order_id=o.order_id
group by c.customer_state;
```

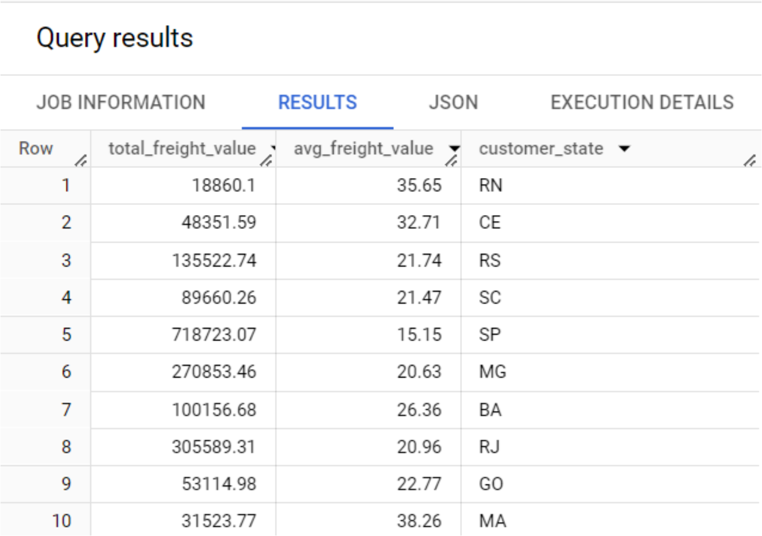


Row	total_price	avg_price	customer_state
1	156453.53	148.3	MT
2	119648.22	145.2	MA
3	80314.81	180.89	AL
4	5202955.05	109.65	SP
5	1585308.03	120.75	MG
6	262788.03	145.51	PE
7	1824092.67	125.12	RJ
8	302603.94	125.77	DF
9	750304.02	120.34	RS
10	58920.85	153.04	SE

- ➢ The average buying price for the state ‘SP’ is comparatively less as compare to all other states and also the total buying price of that state is also maximum.
- ➢ More sales and offers should be done in other state to increase the sales by advertising and spreading awareness about the benefits of online shopping.

3.Calculate the Total & Average value of order freight for each state.
Solution: -

```
select round(sum(ot.freight_value),2) as total_freight_value,
round(avg(ot.freight_value),2) as avg_freight_value,
c.customer_state
from `CaseStudyTarget.customers` c
join
`CaseStudyTarget.orders` o on c.customer_id=o.customer_id
join
`CaseStudyTarget.order_items` ot on ot.order_id=o.order_id
group by c.customer_state;
```



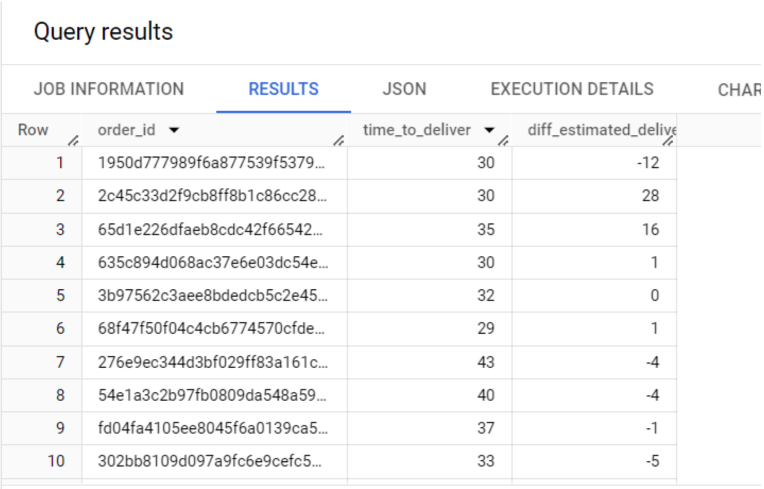
Row	total_freight_value	avg_freight_value	customer_state
1	18860.1	35.65	RN
2	48351.59	32.71	CE
3	135522.74	21.74	RS
4	89660.26	21.47	SC
5	718723.07	15.15	SP
6	270853.46	20.63	MG
7	100156.68	26.36	BA
8	305589.31	20.96	RJ
9	53114.98	22.77	GO
10	31523.77	38.26	MA

5.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:
   - **to_deliver** = order_delivered_customer_date - order_purchase_timestamp
   - **diff_estimated_delivery** = order_estimated_delivery_date - order_delivered_customer_date

Solution: -

```
select order_id,
date_diff(order_delivered_customer_date,order_purchase_timestamp,day) as time_to_deliver,
date_diff(order_estimated_delivery_date,order_delivered_customer_date,day) as
diff_estimated_delivery
from `CaseStudyTarget.orders`;
```



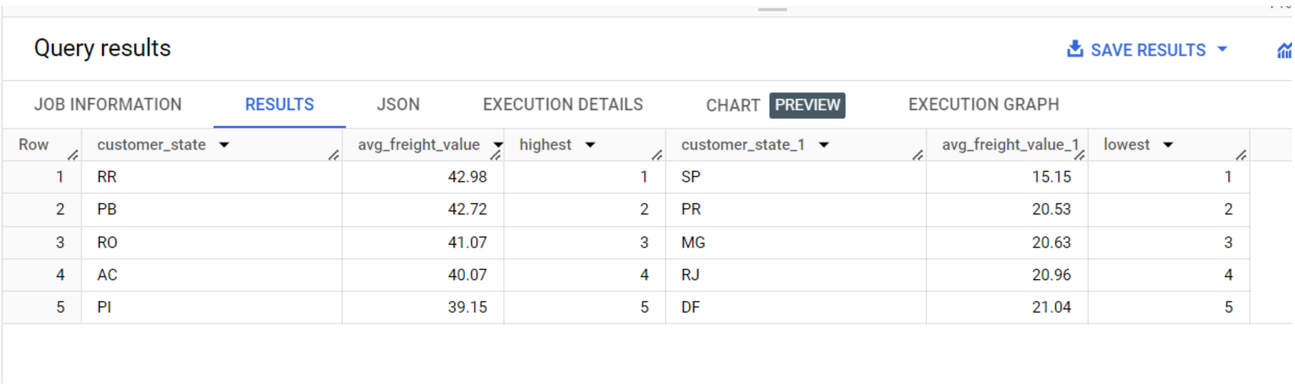
Query results

Row	order_id	time_to_deliver	diff_estimated_delive
1	1950d777989f6a877539f5379...	30	-12
2	2c45c33d2f9cb8ff8b1c86cc28...	30	28
3	65d1e226dfaeb8cdc42f66542...	35	16
4	635c894d068ac37e6e03dc54e...	30	1
5	3b97562c3aee8bdedcb5c2e45...	32	0
6	68f47f50f04c4cb6774570cfde...	29	1
7	276e9ec344d3bf029ff83a161c...	43	-4
8	54e1a3c2b97fb0809da548a59...	40	-4
9	fd04fa4105ee8045f6a0139ca5...	37	-1
10	302bb8109d097a9fc6e9cefc5...	33	-5

Note: -
Time to deliver and diff estimated delivery are in days
Insights: -
The negative sign in diff estimated delivery column indicates that the orders were delivery late after the given estimated delivery date
The positive sign in diff estimated delivery column indicates that the orders were delivery early before the given estimated delivery date

2. Find out the top 5 states with the highest & lowest average freight value.
Solution: -

```
with sidcte as(
select
c.customer_state,
round(avg(ot.freight_value),2) as avg_freight_value
from `CaseStudyTarget.customers` c
join
`CaseStudyTarget.orders` o on c.customer_id=o.customer_id
join
`CaseStudyTarget.order_items` ot on ot.order_id=o.order_id
group by c.customer_state
),
 siddcte as(
  select *,dense_rank() over(order by avg_freight_value desc) as highest,
  dense_rank() over(order by avg_freight_value) as lowest
  from sidcte
)
select a.customer_state,a.avg_freight_value,a.highest,
b.customer_state,b.avg_freight_value,b.lowest
from siddcte a join siddcte b on a.highest=b.lowest
where a.highest<6 and b.lowest<6
order by highest;
```

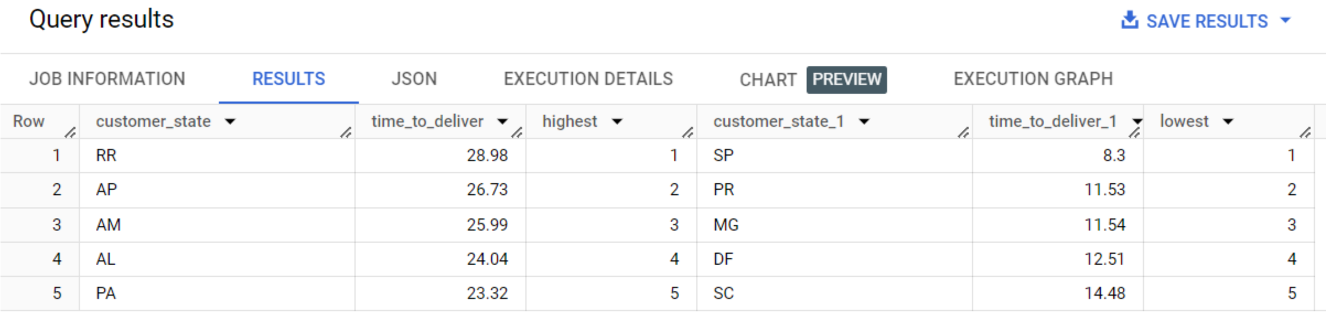


Row	customer_state	avg_freight_value	highest	customer_state_1	avg_freight_value_1	lowest
1	RR	42.98	1	SP	15.15	1
2	PB	42.72	2	PR	20.53	2
3	RO	41.07	3	MG	20.63	3
4	AC	40.07	4	RJ	20.96	4
5	PI	39.15	5	DF	21.04	5

- ➢ These are the top 5 highest and lowest state with their average freight value.

3. Find out the top 5 states with the highest & lowest average delivery time.
Solution: -

```
with sidcte as(
select
c.customer_state,
round(avg(date_diff(o.order_delivered_customer_date,o.order_purchase_timestamp,day)),2) as
time_to_deliver
from `CaseStudyTarget.customers` c
join
`CaseStudyTarget.orders` o on c.customer_id=o.customer_id
group by c.customer_state
),
 siddcte as(
  select *,
  dense_rank() over(order by time_to_deliver desc) as highest,
  dense_rank() over(order by time_to_deliver) as lowest
  from sidcte
)
select a.customer_state,a.time_to_deliver,a.highest,
b.customer_state,b.time_to_deliver,b.lowest
from siddcte a join siddcte b on a.highest=b.lowest
where a.highest<6 and b.lowest<6
order by highest;
```



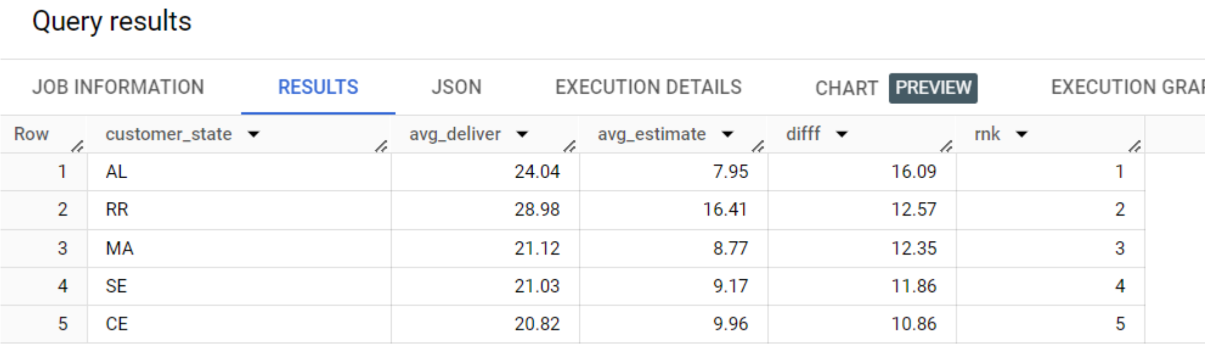
Row	customer_state	time_to_deliver	highest	customer_state_1	time_to_deliver_1	lowest
1	RR	28.98	1	SP	8.3	1
2	AP	26.73	2	PR	11.53	2
3	AM	25.99	3	MG	11.54	3
4	AL	24.04	4	DF	12.51	4
5	PA	23.32	5	SC	14.48	5

- ➢ These are the top 5 highest and lowest state with their average delivery time (Delivery time is in days).

4.Find out the top 5 states where the order delivery is really 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.

Solution: -

```
with sidcte as(
select
c.customer_state,
round(avg(date_diff(o.order_delivered_customer_date,o.order_purchase_timestamp,day)),2) as
avg_deliver,
round(avg(date_diff(o.order_estimated_delivery_date,o.order_delivered_customer_date,day)),2
) as avg_estimate,
from `CaseStudyTarget.customers` c
join
`CaseStudyTarget.orders` o on c.customer_id=o.customer_id
group by c.customer_state
),
siddcte as(
  select *,
  dense_rank() over(order by difff desc) as rnk
  from (
    select *,round(avg_deliver-avg_estimate,2) as difff
    from sidcte
  )
)
select * from siddcte
where rnk<6
order by rnk;
```


Query results

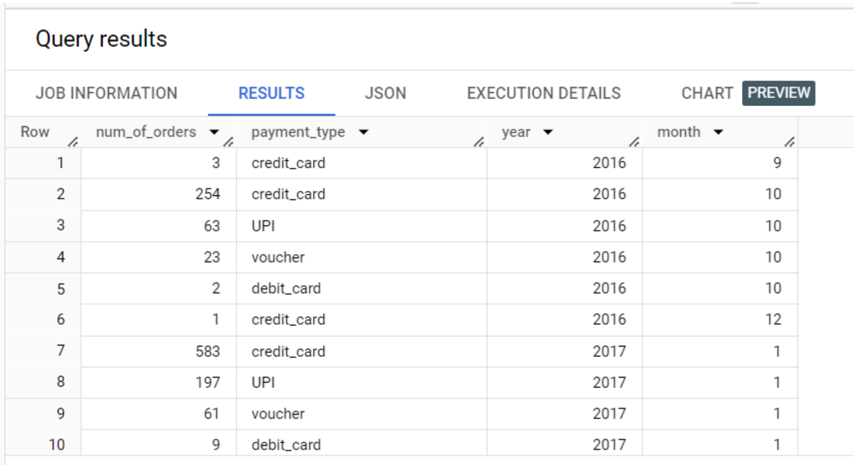
Row	customer_state	avg_deliver	avg_estimate	difff	rnk
1	AL	24.04	7.95	16.09	1
2	RR	28.98	16.41	12.57	2
3	MA	21.12	8.77	12.35	3
4	SE	21.03	9.17	11.86	4
5	CE	20.82	9.96	10.86	5

- These are the top 5 state which has delivery time which is less than estimated delivery time.
- The greater difference between avg estimate and avg deliver denotes that the delivery has been done faster.

6.Analysis based on the payments:

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

```
select count(o.order_id) as num_of_orders,p.payment_type,
extract(year from o.order_purchase_timestamp) as year,
extract(month from o.order_purchase_timestamp) as month
from `CaseStudyTarget.orders` o join `CaseStudyTarget.payments` p using(order_id)
group by 2,3,4
order by year,month;
```

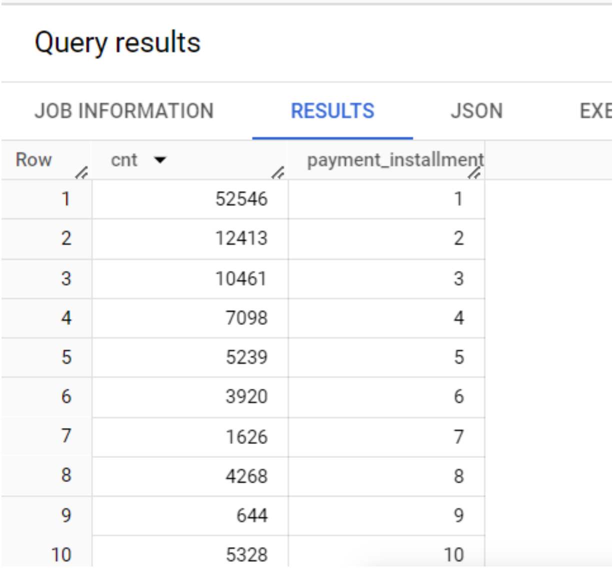


Row	num_of_orders	payment_type	year	month
1	3	credit_card	2016	9
2	254	credit_card	2016	10
3	63	UPI	2016	10
4	23	voucher	2016	10
5	2	debit_card	2016	10
6	1	credit_card	2016	12
7	583	credit_card	2017	1
8	197	UPI	2017	1
9	61	voucher	2017	1
10	9	debit_card	2017	1

➢ The Number of payments done by credit card are more.

2.Find the no. of orders placed on the basis of the payment instalments that have been paid.
Solution: -

```
select count(order_id) as cnt, payment_installments
from `CaseStudyTarget.payments`
group by 2
having payment_installments>0
```



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

➢ The count gradually decreases over payment instalments.

Overall Analysis:

- The data is given of a period of Sep 2016 to Aug 2018
- The month on Sales are increasing over the years from 2016 to 2018.
  Major part of Sales is from the state 'SP'
- Approx. 41% of the customer are from state 'SP' which isn't good sign for the seller in other states and company too. More advertising should be done in other states and new offers on the product should be released frequently.
- The state 'SP' has the lowest freight value which means the travelling cost of product is less in return the profit generated by the state increases, whereas the state 'RR' has highest freight value, so the good network should be established to reduce the travelling cost.
- 'AL' is the state where the delivery time is less. The orders are delivery quickly much before their estimated delivery date.
- Most of the Brazilian Customers do their payment using Credit Card. More rewards can be introduced on credit card and UPI to increase the Sales.