Country -Brazil

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

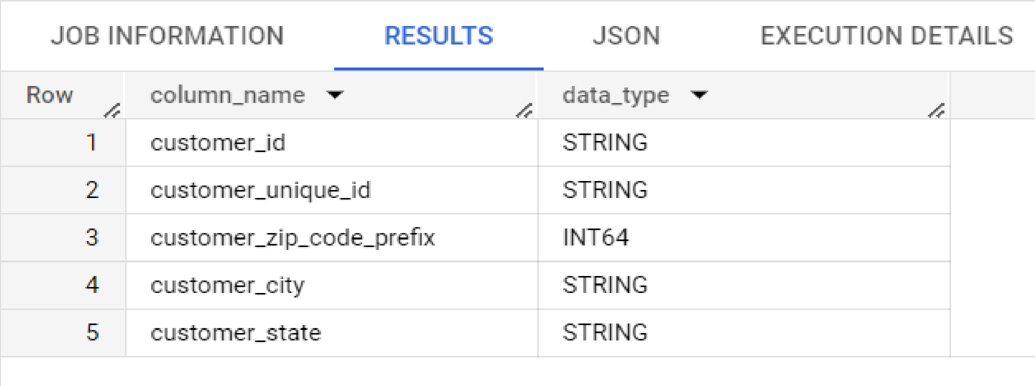
1.1 Data type of all columns in the "customers" table.

Query:

```
SELECT column_name, data_type

FROM sclar_case.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

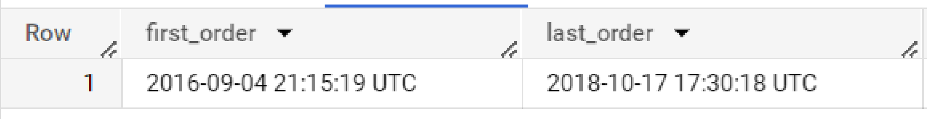
Insights: Most of the columns are of String datatype and only one column is of integer type.

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

Query:

```
SELECT
MIN(order_purchase_timestamp) as first_order,
MAX(order_purchase_timestamp) as last_order
FROM `sclar_case.orders`;
```

Result:



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

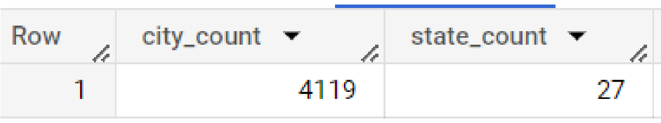
Insights: First order came on 4rt of September 2016 and last order was on 17th October 2018 which is approximately 2 years of time period.

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

Query:

```
select count(distinct customer_city) as city_count, count(distinct customer_state)
as state_count from `sclar_case.customers` c
join `sclar_case.orders` o on o.customer_id=c.customer_id;
```

Result:



Row	city_count	state_count
1	4119	27

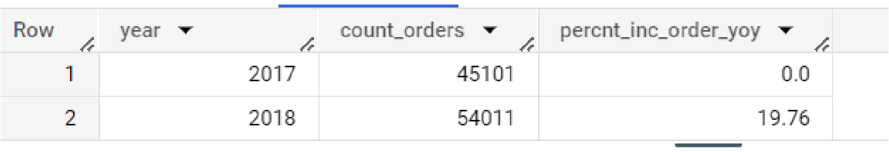
Insights: There are customers from 4119 different cities that are present in 27 different states across Brazil.

2 In-depth Exploration:

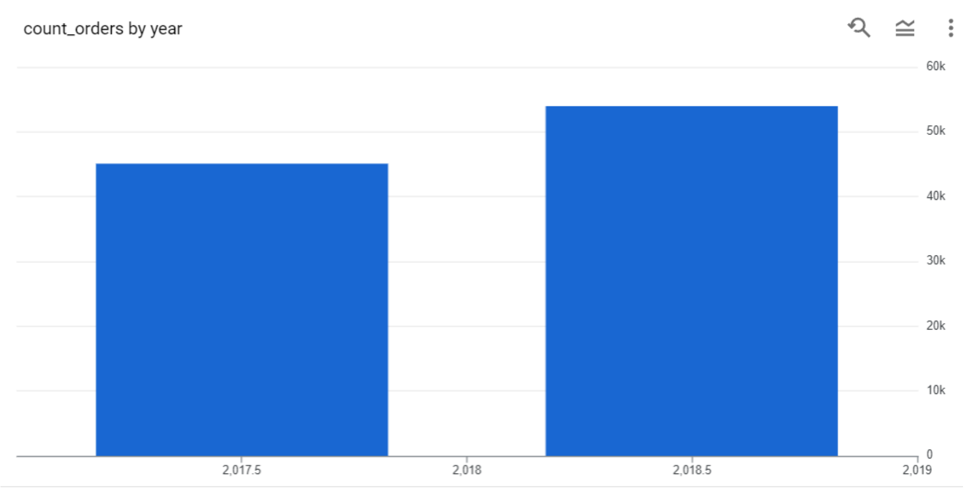
2.1 Is there a growing trend in the no. of orders placed over the past years?

Query:

```
SELECT t.year, t.count_orders,
IFNULL(ROUND(((count_orders-LAG(count_orders) OVER(ORDER BY
t.year))/LAG(count_orders) OVER(ORDER BY t.year))*100,2),0) AS
percnt_inc_order_yoy
FROM
(SELECT EXTRACT(year FROM order_purchase_timestamp) AS year, COUNT(order_id) AS
count_orders FROM `sclar_case.orders`
WHERE EXTRACT(year FROM order_purchase_timestamp)!=2016
GROUP BY EXTRACT(year FROM order_purchase_timestamp)) t;
```



Row	year	count_orders	percnt_inc_order_yoy
1	2017	45101	0.0
2	2018	54011	19.76



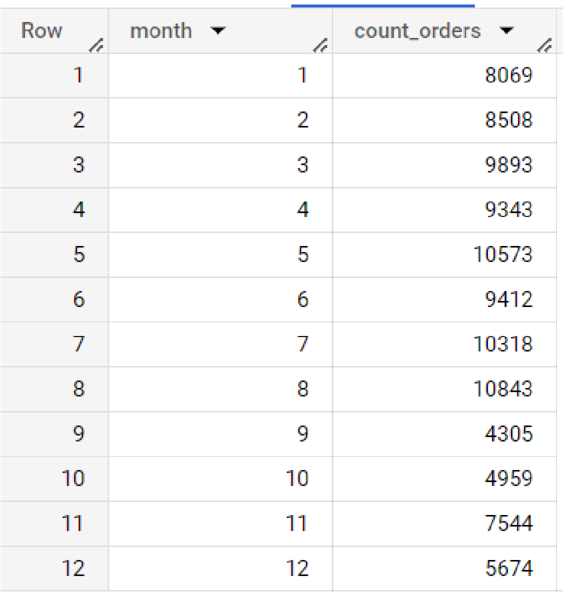
Insights: Yes, there is a growing trend year on year, number of orders in 2018 is greater than that of 2017 even after we have considered 2018 orders up to mid-October only and completely neglecting 2016 orders since market had just started in 2016 September.

Also we can see that the percentage increase of the orders year on year is about 20%.

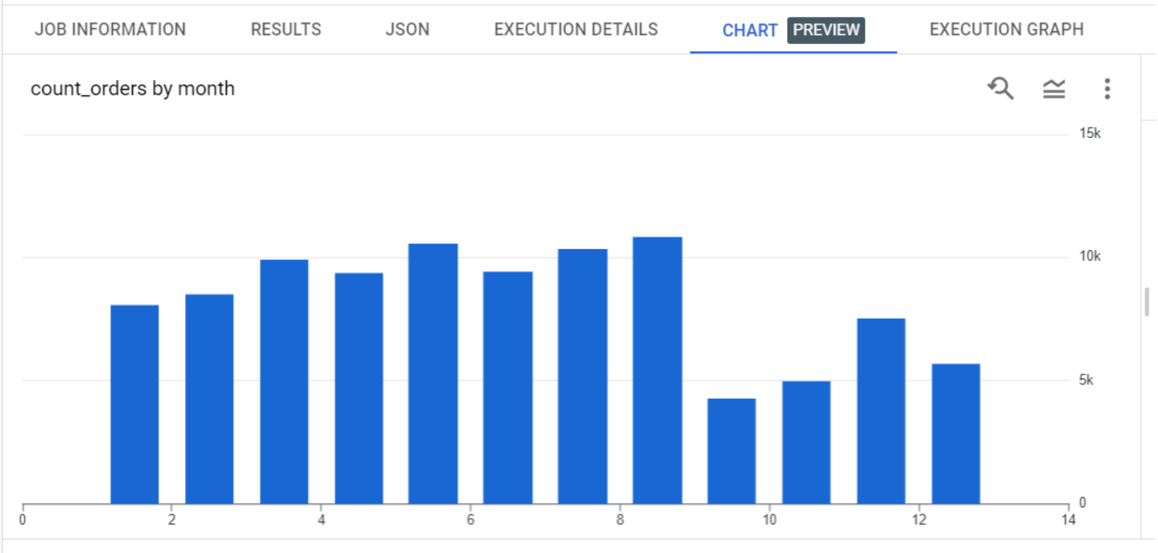
2.2 Can we see some kind of monthly seasonality in terms of the no. of orders being placed?

Query:

```
SELECT t.month, t.count_orders FROM
(SELECT EXTRACT(month FROM order_purchase_timestamp) AS month, COUNT(order_id) AS
count_orders FROM `sclar_case.orders`
GROUP BY EXTRACT(month FROM order_purchase_timestamp)) t
order by t.month;
```



Row	month	count_orders
1	1	8069
2	2	8508
3	3	9893
4	4	9343
5	5	10573
6	6	9412
7	7	10318
8	8	10843
9	9	4305
10	10	4959
11	11	7544
12	12	5674



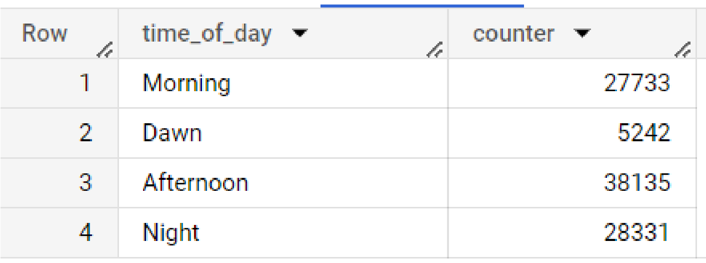
Insights: We can see that most of the orders are coming in first 3 quarters of the year and very few orders are placed in last quarter of the year with least orders in the month of September.

2.3 During what time of the day, do the Brazilian customers mostly place their orders? (Dawn, Morning, Afternoon or Night)

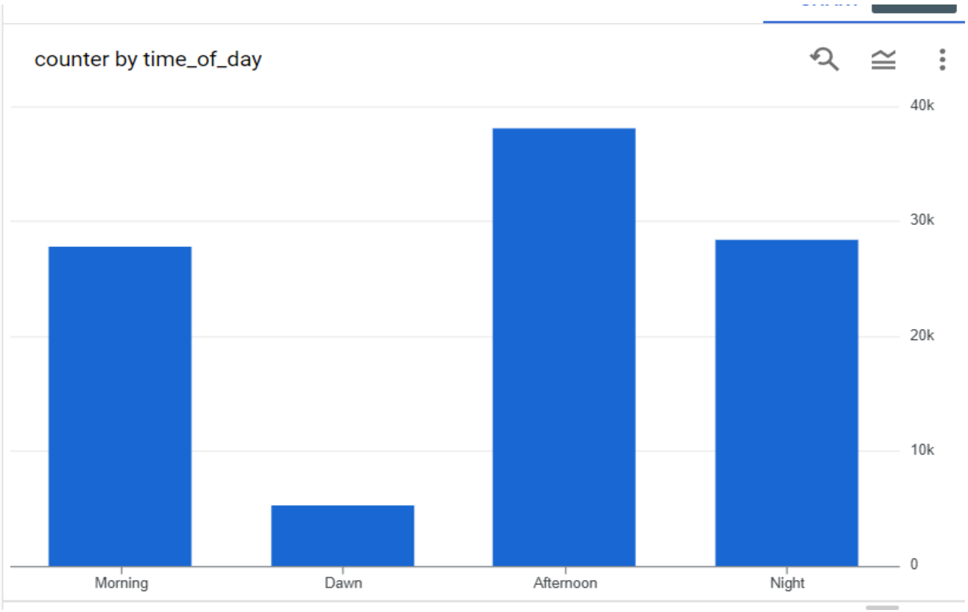
0-6 hrs: Dawn
7-12 hrs: Mornings
13-18 hrs: Afternoon
19-23 hrs: Night

Query:

```
select  t.time_of_day, count(*) as counter from
(select case
when extract(hour from order_purchase_timestamp) between 0 and 6 then 'Dawn'
when extract(hour from order_purchase_timestamp) between 6 and 12 then 'Morning'
when extract(hour from order_purchase_timestamp) between 12 and 18 then
'Afternoon'
else 'Night'
end as time_of_day
from `sclar_case.orders`) t
group by t.time_of_day;
```



Row	time_of_day	counter
1	Morning	27733
2	Dawn	5242
3	Afternoon	38135
4	Night	28331



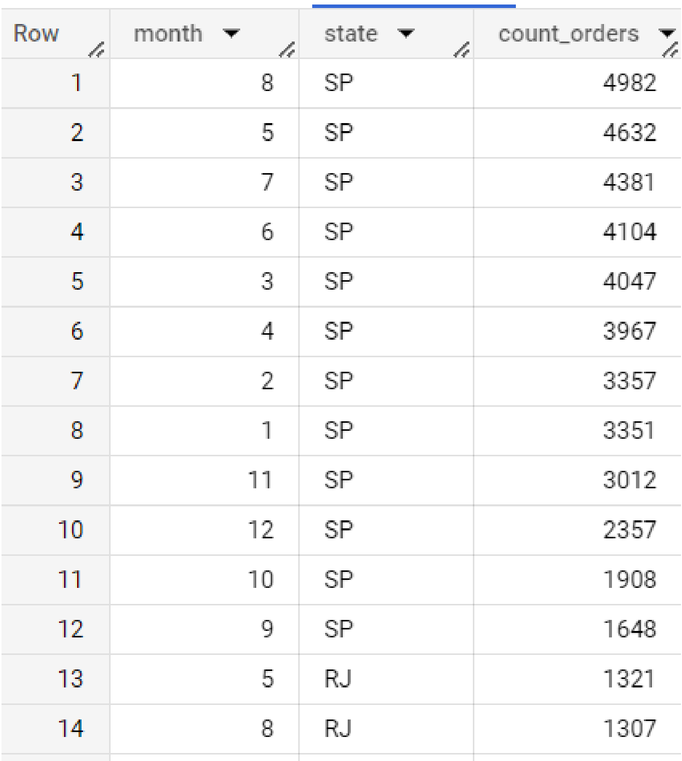
Insights: Most of the orders are placed during afternoon and least orders are coming from dawn period i.e. mostly customers find the leisure/free time to shop online during afternoon or night period.

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

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

Query:

```
select t.month, t.state, t.count_orders from
(select extract(month from order_purchase_timestamp) as month, customer_state as
state ,count(distinct order_id) as count_orders from `sclar_case.orders` o join
`sclar_case.customers` c on c.customer_id=o.customer_id
group by extract(month from order_purchase_timestamp), customer_state) t
order by t.count_orders desc;
```



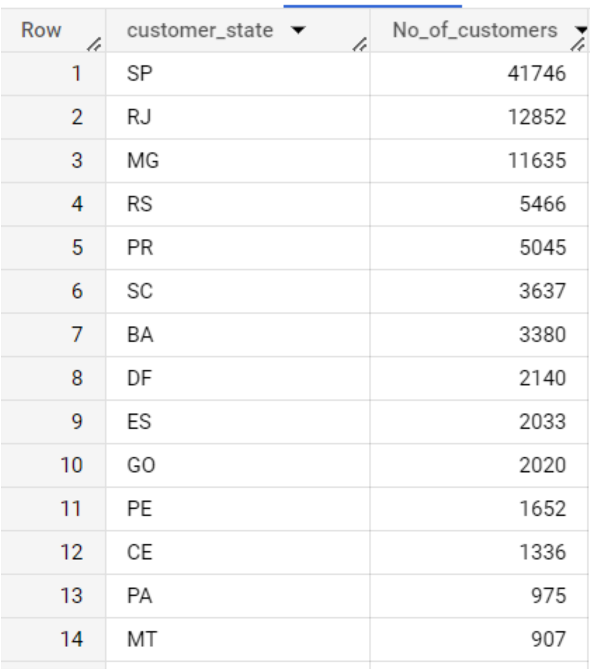
Row	month	state	count_orders
1	8	SP	4982
2	5	SP	4632
3	7	SP	4381
4	6	SP	4104
5	3	SP	4047
6	4	SP	3967
7	2	SP	3357
8	1	SP	3351
9	11	SP	3012
10	12	SP	2357
11	10	SP	1908
12	9	SP	1648
13	5	RJ	1321
14	8	RJ	1307

Insights: We can see that maximum orders are coming from SP state throughout the year.

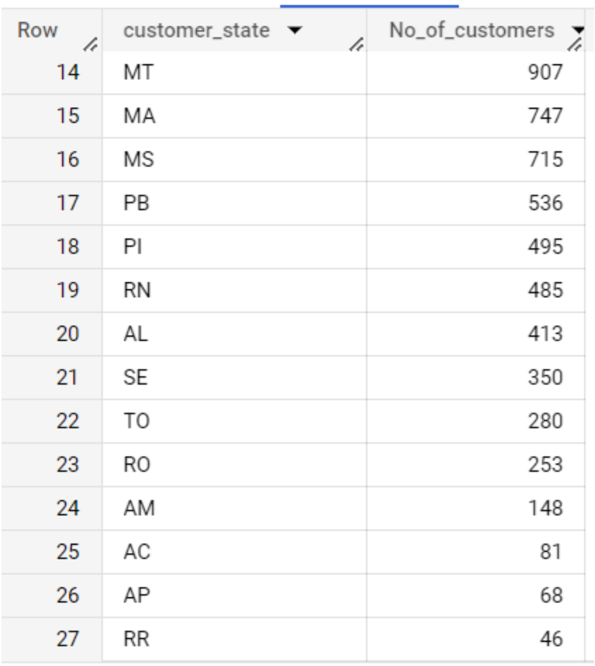
3.2 How are the customers distributed across all the states?

Query:

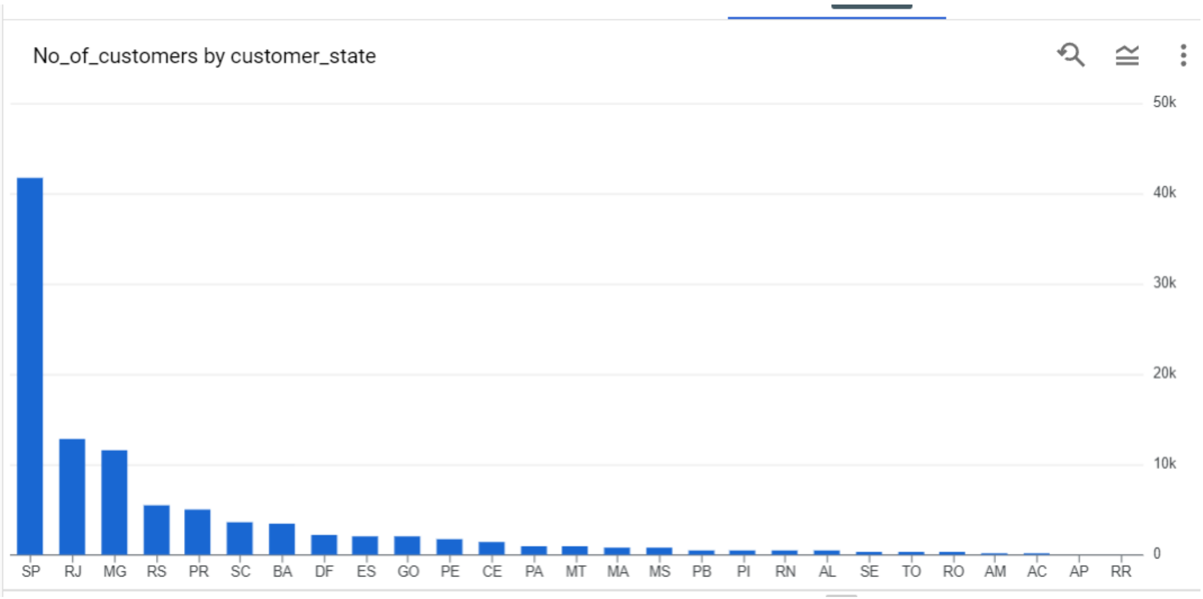
```
select customer_state, count(distinct c.customer_id) as No_of_customers from
`sclar_case.customers` c
join `sclar_case.orders` o on o.customer_id=c.customer_id
group by customer_state
order by No_of_customers desc;
```



Row	customer_state	No_of_customers
1	SP	41746
2	RJ	12852
3	MG	11635
4	RS	5466
5	PR	5045
6	SC	3637
7	BA	3380
8	DF	2140
9	ES	2033
10	GO	2020
11	PE	1652
12	CE	1336
13	PA	975
14	MT	907



Row	customer_state	No_of_customers
14	MT	907
15	MA	747
16	MS	715
17	PB	536
18	PI	495
19	RN	485
20	AL	413
21	SE	350
22	TO	280
23	RO	253
24	AM	148
25	AC	81
26	AP	68
27	RR	46



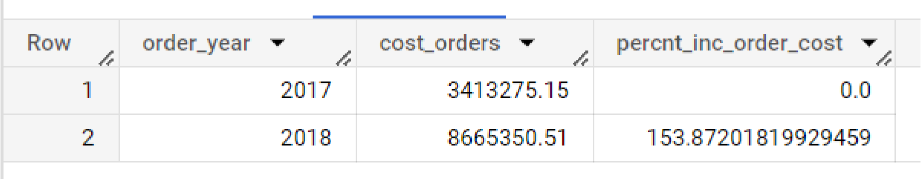
Insights: Maximum customers are from state SP and least are from RR. About 42% of customers are from a single state SP.

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

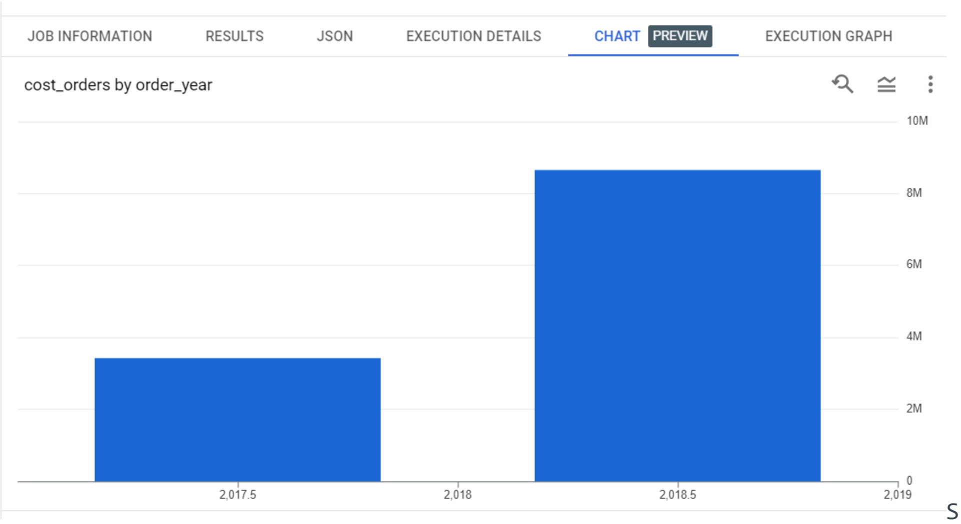
4.1 Get the % increase in the cost of orders from year 2017 to 2018 (include months between Jan to Aug only).

Query:

```
select t.order_year, t.cost_orders,
ifnull(((t.cost_orders-lag(t.cost_orders) over(order by
t.order_year))/(lag(t.cost_orders) over(order by t.order_year)))*100,0) as
percnt_inc_order_cost
from
(select extract(year from order_delivered_carrier_date) as order_year,
round(sum(payment_value),2) as cost_orders
from `sclar_case.orders` o
join `sclar_case.payments` p on o.order_id=p.order_id
where extract(year from o.order_delivered_carrier_date)!=2016 and
(o.order_delivered_carrier_date between '2017-01-01' and '2017-08-31')
or (o.order_delivered_carrier_date between '2018-01-01' and '2018-08-31')
group by extract(year from o.order_delivered_carrier_date)) t
order by t.order_year;
```



Row	order_year	cost_orders	percnt_inc_order_cost
1	2017	3413275.15	0.0
2	2018	8665350.51	153.87201819929459



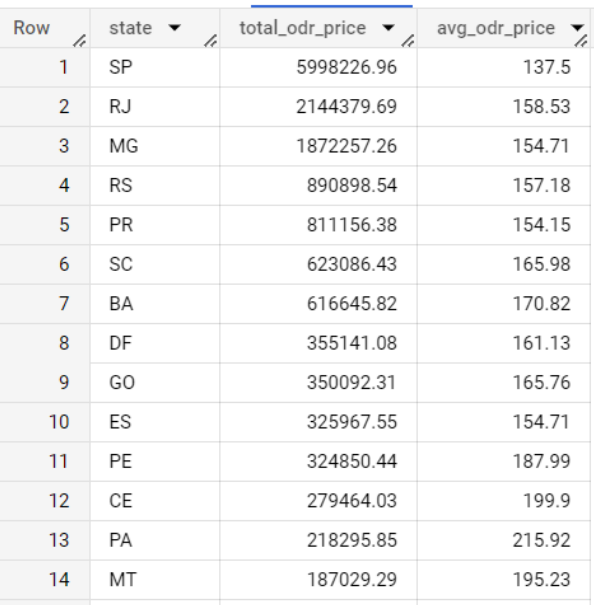
S

Insights: Order cost has increased by whooping 154% from 2017 to 2018 for period between January to August.

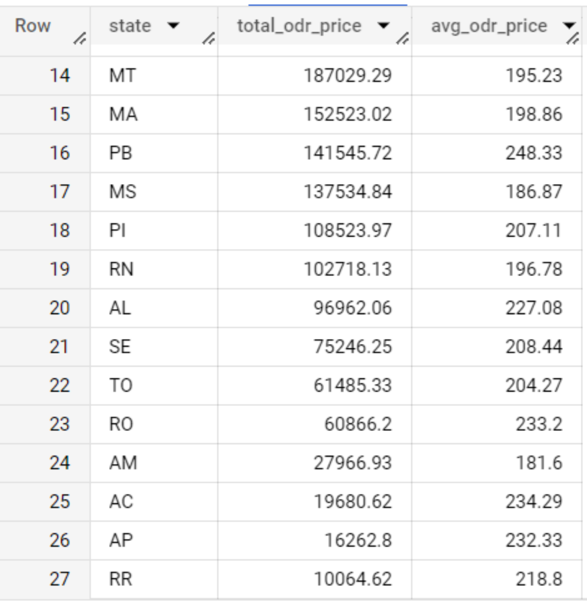
4.2 Calculate the Total & Average value of order price for each state.

Query:

```
select customer_state as state, round(sum(payment_value),2) as total_odr_price,
round(sum(payment_value)/count(payment_value),2) as avg_odr_price
from `sclar_case.payments` p
join `sclar_case.orders` o on o.order_id=p.order_id join `sclar_case.customers` c
on o.customer_id=c.customer_id
group by customer_state
order by total_odr_price desc;
```



Row	state	total_odr_price	avg_odr_price
1	SP	5998226.96	137.5
2	RJ	2144379.69	158.53
3	MG	1872257.26	154.71
4	RS	890898.54	157.18
5	PR	811156.38	154.15
6	SC	623086.43	165.98
7	BA	616645.82	170.82
8	DF	355141.08	161.13
9	GO	350092.31	165.76
10	ES	325967.55	154.71
11	PE	324850.44	187.99
12	CE	279464.03	199.9
13	PA	218295.85	215.92
14	MT	187029.29	195.23



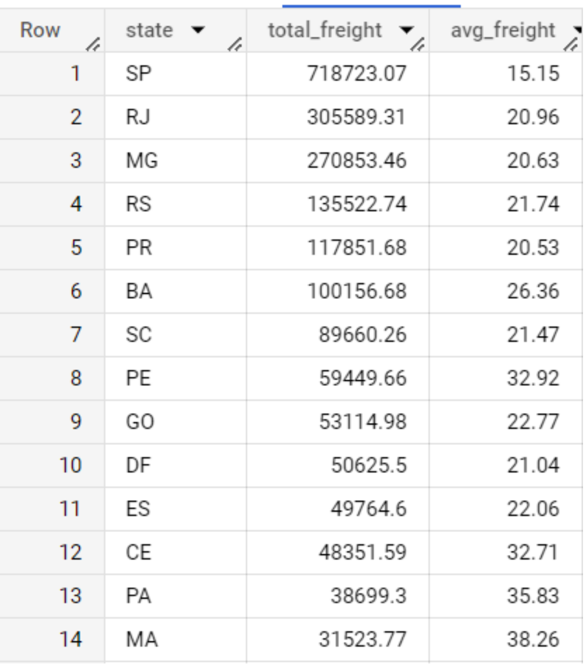
Row	state	total_odr_price	avg_odr_price
14	MT	187029.29	195.23
15	MA	152523.02	198.86
16	PB	141545.72	248.33
17	MS	137534.84	186.87
18	PI	108523.97	207.11
19	RN	102718.13	196.78
20	AL	96962.06	227.08
21	SE	75246.25	208.44
22	TO	61485.33	204.27
23	RO	60866.2	233.2
24	AM	27966.93	181.6
25	AC	19680.62	234.29
26	AP	16262.8	232.33
27	RR	10064.62	218.8

Insights: Maximum total order price is from ‘SP’ state but the highest average order price is of ‘PB’ state.

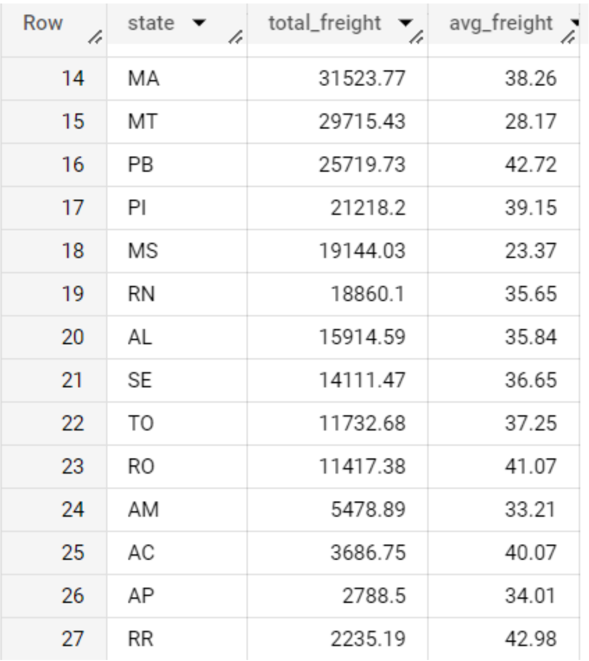
4.3 Calculate the Total & Average value of order freight for each state.

Query:

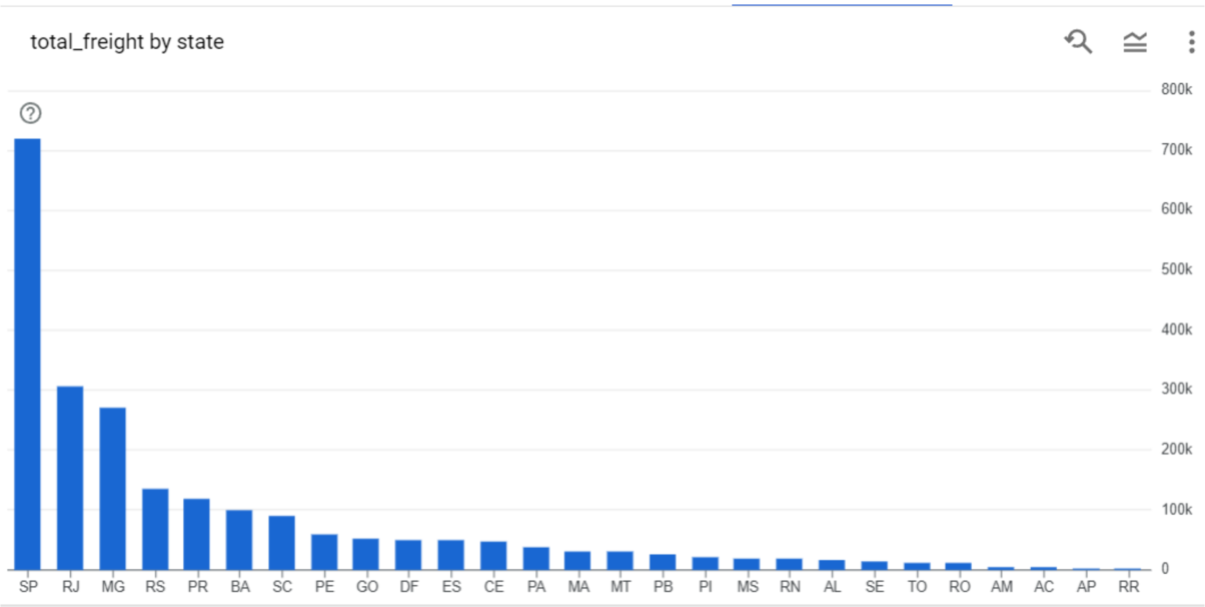
```
select customer_state as state, round(sum(freight_value),2) as total_freight,
round(sum(freight_value)/count(freight_value),2) as avg_freight
from `sclar_case.order_items` oi
join `sclar_case.orders` o on o.order_id=oi.order_id join `sclar_case.customers` c
on c.customer_id=o.customer_id
group by customer_state
order by total_freight desc;
```



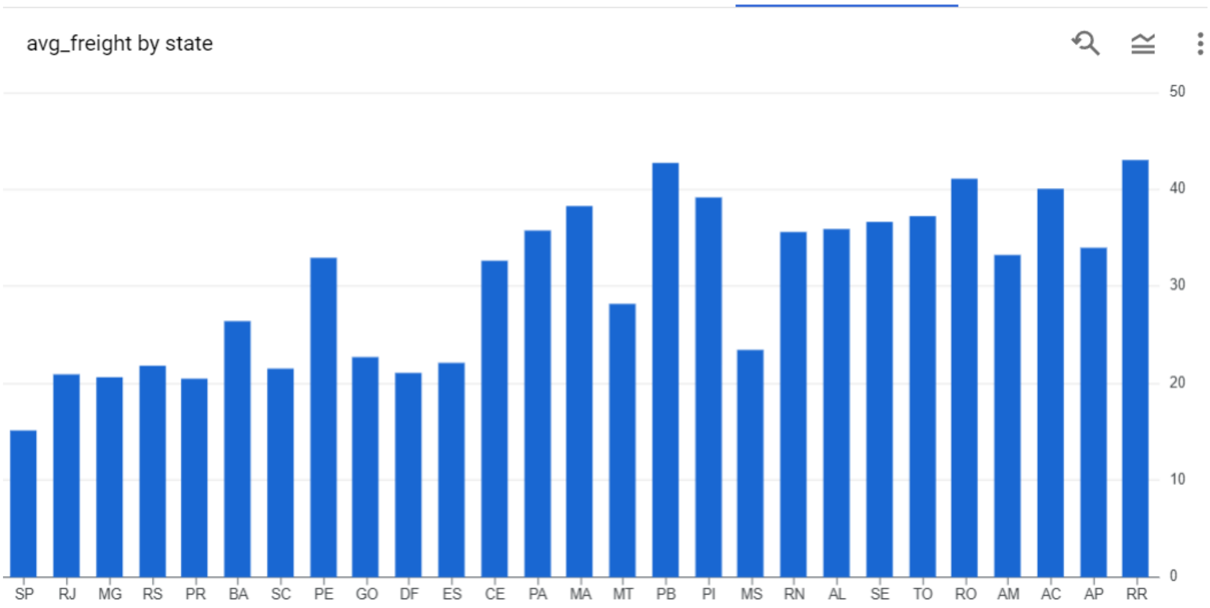
Row	state	total_freight	avg_freight
1	SP	718723.07	15.15
2	RJ	305589.31	20.96
3	MG	270853.46	20.63
4	RS	135522.74	21.74
5	PR	117851.68	20.53
6	BA	100156.68	26.36
7	SC	89660.26	21.47
8	PE	59449.66	32.92
9	GO	53114.98	22.77
10	DF	50625.5	21.04
11	ES	49764.6	22.06
12	CE	48351.59	32.71
13	PA	38699.3	35.83
14	MA	31523.77	38.26



Row	state	total_freight	avg_freight
14	MA	31523.77	38.26
15	MT	29715.43	28.17
16	PB	25719.73	42.72
17	PI	21218.2	39.15
18	MS	19144.03	23.37
19	RN	18860.1	35.65
20	AL	15914.59	35.84
21	SE	14111.47	36.65
22	TO	11732.68	37.25
23	RO	11417.38	41.07
24	AM	5478.89	33.21
25	AC	3686.75	40.07
26	AP	2788.5	34.01
27	RR	2235.19	42.98



total_freight by state



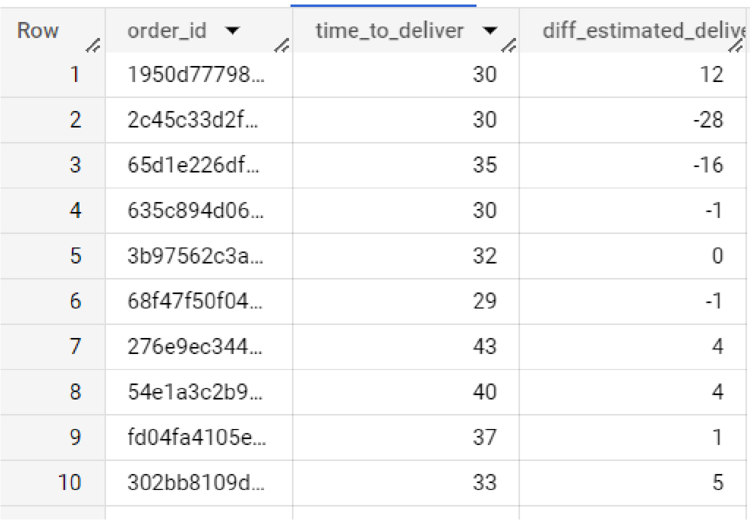
Insights: Here we can see that the state with highest total freight is having minimum average freight value which is expected because maximum number of customers are from these states which can be seen from 3.2. So it is good that to have minimum freight charges where it could benefit maximum customers. Company can in turn benefit by more number of orders in this case.

5 Analysis based on sales, freight and delivery time.

5.1 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.

Query:

```
SELECT order_id, TIMESTAMP_DIFF(order_delivered_customer_date,
order_purchase_timestamp, DAY) AS time_to_deliver,
TIMESTAMP_DIFF(order_delivered_customer_date,order_estimated_delivery_date, DAY)
AS diff_estimated_delivery
FROM `sclar_case.orders`
```



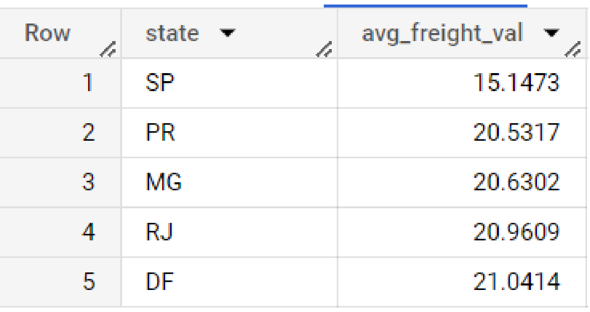
Row	order_id	time_to_deliver	diff_estimated_delive
1	1950d77798...	30	12
2	2c45c33d2f...	30	-28
3	65d1e226df...	35	-16
4	635c894d06...	30	-1
5	3b97562c3a...	32	0
6	68f47f50f04...	29	-1
7	276e9ec344...	43	4
8	54e1a3c2b9...	40	4
9	fd04fa4105e...	37	1
10	302bb8109d...	33	5

Insights: Time to deliver is mostly between 30 days to 50 days which is greater than expected. Difference in estimated and actual delivery time has much greater variance in values, negative values in diff_estimated_delivery indicate that the product was delivered before the expected date which is good. But we can also see many positive values here which can damage the goodwill of the company.

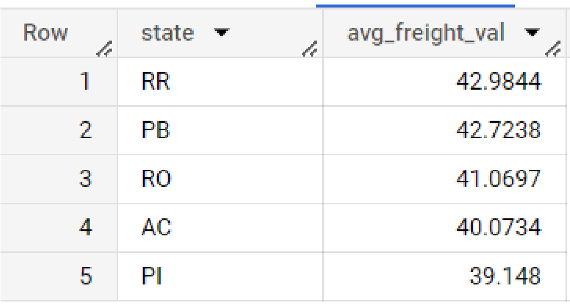
5.2 Find out the top 5 states with the highest & lowest average freight value.

Query:

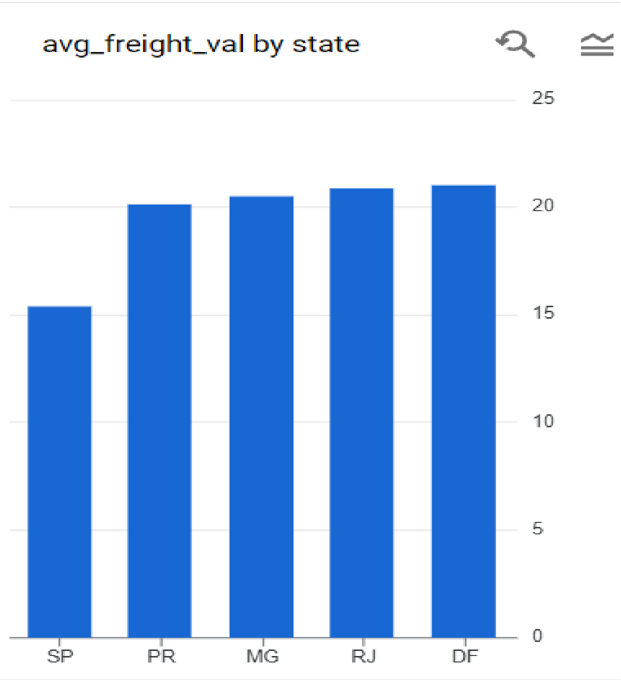
```
select customer_state as state, round(sum(freight_value)/count(freight_value),4)
as avg_freight_val from `sclar_case.order_items` oi
join `sclar_case.orders` o on o.order_id=oi.order_id join `sclar_case.customers` c
on c.customer_id=o.customer_id
group by customer_state
order by avg_freight_val
limit 5;
```

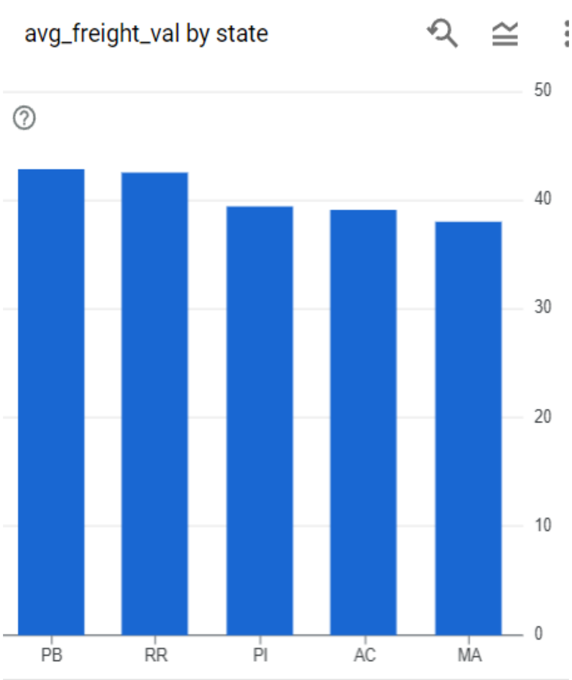


Row	state	avg_freight_val
1	SP	15.1473
2	PR	20.5317
3	MG	20.6302
4	RJ	20.9609
5	DF	21.0414



Row	state	avg_freight_val
1	RR	42.9844
2	PB	42.7238
3	RO	41.0697
4	AC	40.0734
5	PI	39.148



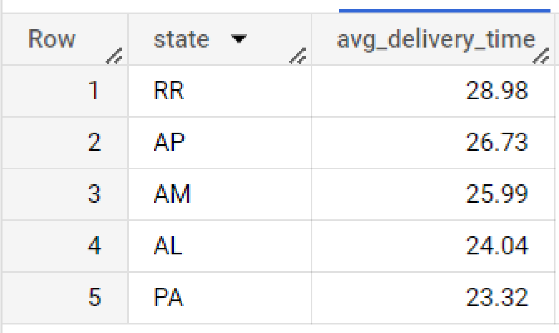


Insights: Top 5 states with maximum average freight value are PB, RR, PI, AC, MA and bottom 5 average freight values are of states SP, PR, MG, RJ, DF with maximum value of 42.8 and minimum value of 15.4 out of total 27 states.
Comparing with 3.2, we can see that the average freight values are less in the states where the no of customers are more which is good.

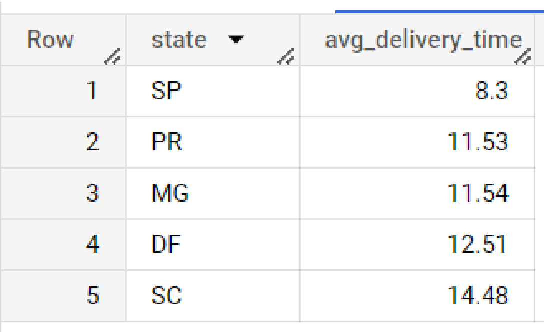
5.3 Find out the top 5 states with the highest & lowest average delivery time.

Query:

```
select t.state, round(sum(t.time_to_deliver)/count(t.time_to_deliver),2) as
avg_delivery_time from
(select customer_state as state, TIMESTAMP_DIFF(order_delivered_customer_date,
order_purchase_timestamp, DAY) AS time_to_deliver,
from `sclar_case.orders` o join `sclar_case.customers` c on
o.customer_id=c.customer_id) t
group by t.state
order by avg_delivery_time desc
limit 5;
```



Row	state	avg_delivery_time
1	RR	28.98
2	AP	26.73
3	AM	25.99
4	AL	24.04
5	PA	23.32



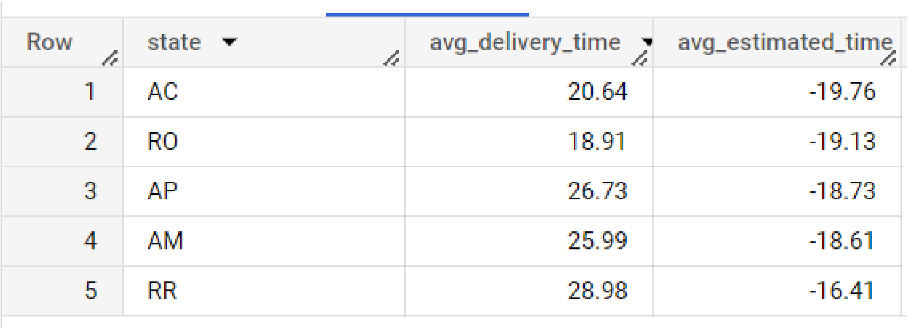
Row	state	avg_delivery_time
1	SP	8.3
2	PR	11.53
3	MG	11.54
4	DF	12.51
5	SC	14.48

Insights: We have SP, PR, MG, DF and SC as top 5 states in terms of average delivery time having least delivery time which is good because most of the customers are from these states. And bottom 5 states are AP, AM, RR, AL and PA having maximum average delivery time.

5.4 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.

Query:

```
select t.state, round(sum(t.time_to_deliver)/count(t.time_to_deliver),2) as
avg_delivery_time,
round(sum(t.diff_estimated_delivery)/count(t.diff_estimated_delivery),2) as
avg_estimated_time
from
(select customer_state as state, TIMESTAMP_DIFF(order_delivered_customer_date,
order_purchase_timestamp, DAY) AS time_to_deliver,
TIMESTAMP_DIFF(order_delivered_customer_date,order_estimated_delivery_date, DAY)
AS diff_estimated_delivery
from `sclar_case.orders` o join `sclar_case.customers` c on
o.customer_id=c.customer_id) t
group by t.state
order by avg_estimated_time
limit 5;
```



Row	state	avg_delivery_time	avg_estimated_time
1	AC	20.64	-19.76
2	RO	18.91	-19.13
3	AP	26.73	-18.73
4	AM	25.99	-18.61
5	RR	28.98	-16.41

Insights: The top 5 states having delivery time really fast as compared to estimated time are RR, AM, RO, AC and AP. Negative values in average estimated time indicated the order was delivered before estimated date. More is the magnitude of value in negative, quicker is the order delivered.

---

6 Analysis based on the payments:

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

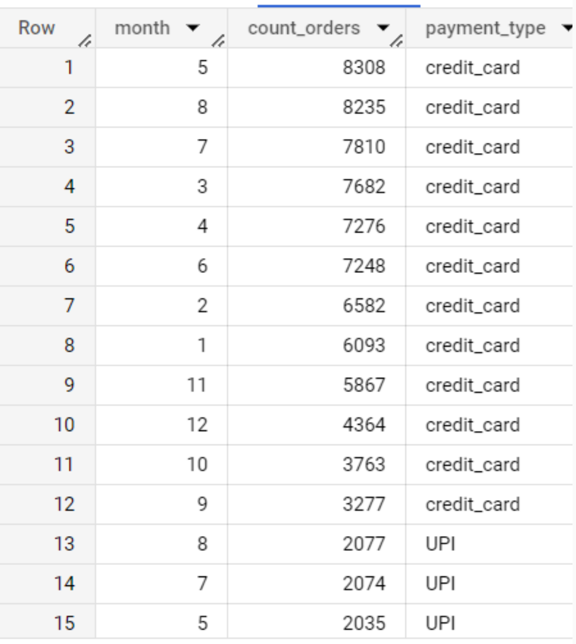
Query:

```
select EXTRACT(month FROM o.order_purchase_timestamp) AS month, count(distinct
o.order_id) as count_orders, payment_type
```

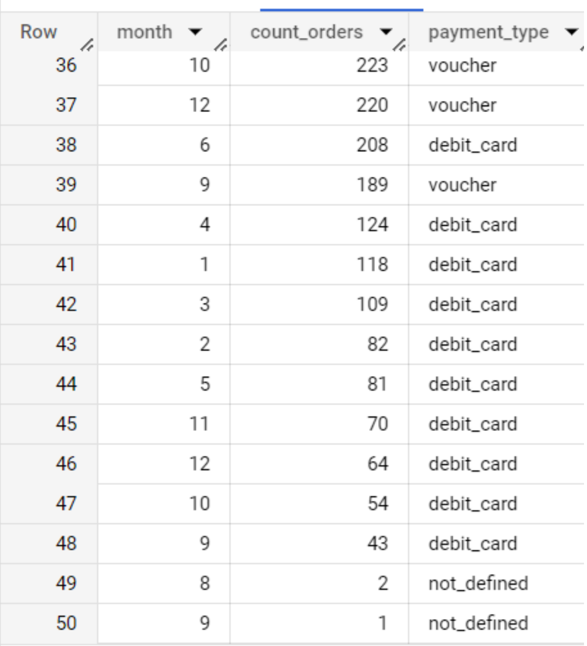
```
from `sclar_case.payments` p join `sclar_case.orders` o on o.order_id=p.order_id
where payment_type='credit_card'
group by EXTRACT(month FROM o.order_purchase_timestamp), payment_type
union all
select EXTRACT(month FROM o.order_purchase_timestamp) AS month, count(distinct
o.order_id) as count_orders, payment_type
from `sclar_case.payments` p join `sclar_case.orders` o on o.order_id=p.order_id
where payment_type='voucher'
group by EXTRACT(month FROM o.order_purchase_timestamp), payment_type
union all
select EXTRACT(month FROM o.order_purchase_timestamp) AS month, count(distinct
o.order_id) as count_orders, payment_type
from `sclar_case.payments` p join `sclar_case.orders` o on o.order_id=p.order_id
where payment_type='not_defined'
group by EXTRACT(month FROM o.order_purchase_timestamp), payment_type
union all
select EXTRACT(month FROM o.order_purchase_timestamp) AS month, count(distinct
o.order_id) as count_orders, payment_type
from `sclar_case.payments` p join `sclar_case.orders` o on o.order_id=p.order_id
where payment_type='debit_card'
group by EXTRACT(month FROM o.order_purchase_timestamp), payment_type
union all
select EXTRACT(month FROM o.order_purchase_timestamp) AS month, count(distinct
o.order_id) as count_orders, payment_type
from `sclar_case.payments` p join `sclar_case.orders` o on o.order_id=p.order_id
where payment_type='UPI'
group by EXTRACT(month FROM o.order_purchase_timestamp), payment_type
order by t.count_orders desc;
```

OR

```
select t.month, t.count_orders, t.payment_type from
(select EXTRACT(month FROM o.order_purchase_timestamp) AS month, count(distinct
o.order_id) as count_orders, payment_type
from `sclar_case.payments` p join `sclar_case.orders` o on o.order_id=p.order_id
group by EXTRACT(month FROM o.order_purchase_timestamp), payment_type) t
order by t.count_orders desc;
```



Row	month	count_orders	payment_type
1	5	8308	credit_card
2	8	8235	credit_card
3	7	7810	credit_card
4	3	7682	credit_card
5	4	7276	credit_card
6	6	7248	credit_card
7	2	6582	credit_card
8	1	6093	credit_card
9	11	5867	credit_card
10	12	4364	credit_card
11	10	3763	credit_card
12	9	3277	credit_card
13	8	2077	UPI
14	7	2074	UPI
15	5	2035	UPI



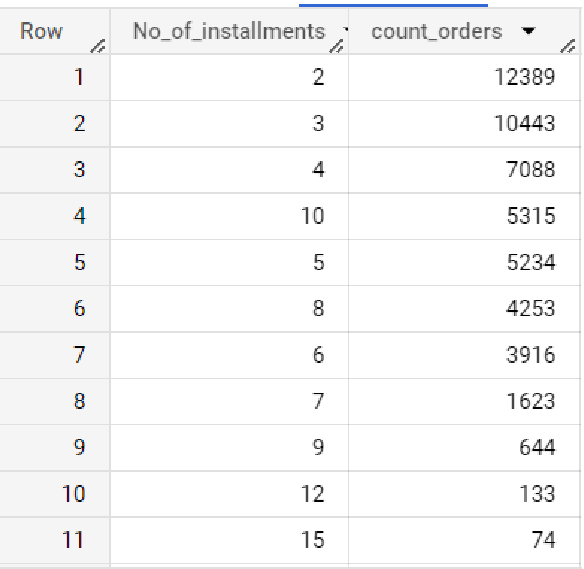
Row	month	count_orders	payment_type
36	10	223	voucher
37	12	220	voucher
38	6	208	debit_card
39	9	189	voucher
40	4	124	debit_card
41	1	118	debit_card
42	3	109	debit_card
43	2	82	debit_card
44	5	81	debit_card
45	11	70	debit_card
46	12	64	debit_card
47	10	54	debit_card
48	9	43	debit_card
49	8	2	not_defined
50	9	1	not_defined

Insights: Most of the orders were paid via credit card and least with debit card.

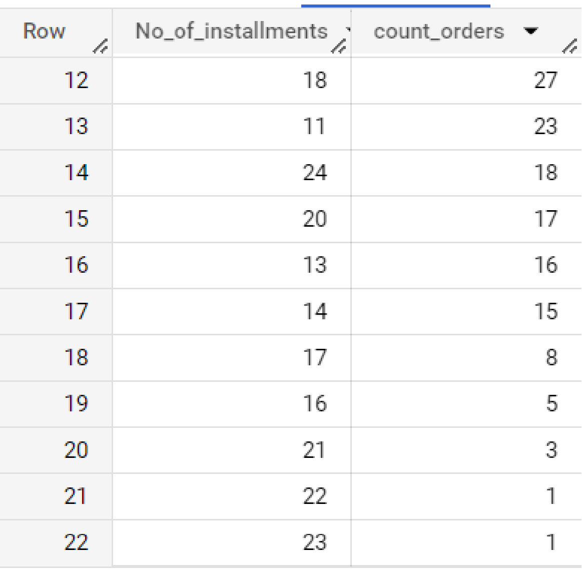
6.2 Find the no. of orders placed on the basis of the payment installments that have been paid.

Query:

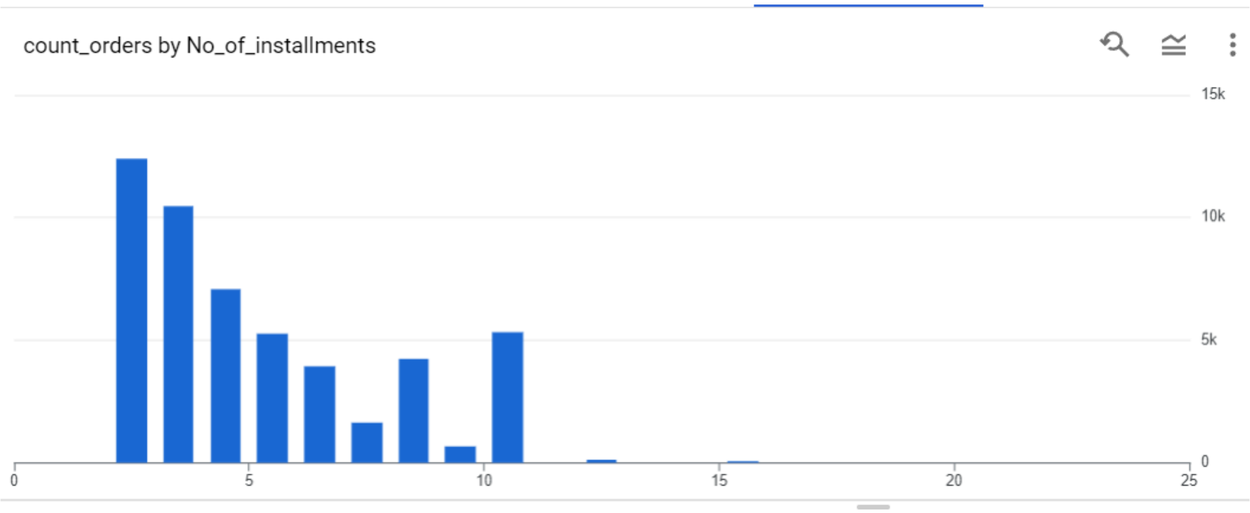
```
select payment_installments as No_of_installments, count(distinct o.order_id) as
count_orders from `sclar_case.payments` p
join `sclar_case.orders` o on p.order_id=o.order_id
WHERE payment_installments>1
group by payment_installments
order by count_orders desc;
```



Row	No_of_installments	count_orders
1	2	12389
2	3	10443
3	4	7088
4	10	5315
5	5	5234
6	8	4253
7	6	3916
8	7	1623
9	9	644
10	12	133
11	15	74



Row	No_of_installments	count_orders
12	18	27
13	11	23
14	24	18
15	20	17
16	13	16
17	14	15
18	17	8
19	16	5
20	21	3
21	22	1
22	23	1



Insights: Most of the orders are paid in single instalments i.e. most of the customers are comfortable paying right after the purchase which is showing good purchase parity of the consumers.

---

7 Actionable Insights & Recommendations

7.1 Since from 2.2 we can conclude that number of orders are considerably low during the last quarter of the year, so company should consider giving some special offers or discounts during that phase to boost up the sales.

7.2 From 2.3 we can conclude that If company wants to promote or advertise about any product or new offer, noon or night time slot would be most effective and cost efficient for the company. Since most of the customer's place orders during this time, so company could reach out to maximum customers with minimum cost of promotion.

7.3 From 3.1 and 3.2 we can conclude that most of the customers are concentrated in one state itself i.e. SP which is not recommended for smooth functioning of any multi-national company in long run. Company is very much dependent on a single state. Company needs to expand services in other states as well. And to expand its customer base company needs to come up with some strategic offers suitable for the local customers in that state.

7.4 From 4.1 we can conclude that order cost has considerably increased by about 150% in just 1 year which is not recommended because such a rapid order cost increase would affect the profit margins of the company.

7.5 From 4.2 and 4.3 we can conclude that the state with maximum orders(SP) is having least average price per order. So company need to focus on boosting the sales of high value products in SP and also expand the customer base in states where company is getting high value orders. Cost of orders and freight charges are high in states where orders placed are less which is obvious but company should focus on expanding the customer base in other states as well by decreasing the freight charges and improving the delivery network of the company which could attract more customers.

7.6 From 5.1 we can conclude that delivery time mostly around 30 to 50 days which needs to be reduced i.e. the warehouses of the company should be more evenly distributed and there should be adequate warehouses in the states where the customers are more. Also company should take steps to deliver the products within estimated delivery date.

7.7 From 6.1 we can conclude that customers of the company are mostly using credit cards for payment and from 6.2 we can say that customers are mostly paying the order price in a single instalment which indicates customers have good purchasing parity and company can also promote some high ticket/high margin products to the customers and boost its sales and margins.