What does 'good' look like?

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

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

```
SELECT column_name, data_type
FROM project-one-415116.target.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

Insight: Understanding the data types of each column in the tables is essential for data manipulation and analysis. It allows us to know the nature of the data stored in each column, facilitating appropriate querying and processing.

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

```
select
min(order_purchase_timestamp),
max(order_purchase_timestamp)
from
target.orders
```

Press Alt+F1 for Accessibility Options.

Query results SAVE RESULTS

JOB INFORMATION | RESULTS | CHART | JSON

Row	f0_	f1_
1	2016-09-04 21:15:19 UTC	2018-10-17 17:30:18 UTC

Insight: Knowing the time range during which orders were placed provides us with valuable information about the duration covered by the dataset. This insight is crucial for understanding the historical scale of our analysis and can help identify any trends or patterns over time

The orders were placed between September 4th 2016 and 17th October 2018

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

```
select
count (distinct customer_city) as `no_of_city`,
count (distinct customer_state) as `no_of_state`
from
target.customers
```

Row	no_of_city	no_of_state
1	4119	27

Insight: By counting the number of unique cities and states from which orders were placed, we gain insight into the geographic distribution of customers. This information can be useful for identifying

regions with high customer engagement and potential areas for targeted marketing or expansion efforts.

Customers have ordered from 27 states and 4119 cities.

2. In-depth Exploration:

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

```
1  select
2  count(order_id) as `number_of_orders`,
3  EXTRACT(YEAR FROM order_purchase_timestamp) as `year`
4  from
5  target.orders
6  group by year
7  order by year desc
8
```

Press Alt+F1 for Accessibility Options.

Query results SAVE RESULTS

JOB INFORMATION | RESULTS | CHART | JSON

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

Insight: Analyzing the trend in the number of orders placed over the past years is essential for understanding the business's growth trajectory. By examining the order volume month by month or year by year, we can identify whether there is a consistent upward trend indicating growth or not. This insight is crucial for strategic planning, resource allocation, and forecasting future business performance.

The available data for years 2016 and 2018 is limited. With the available data we can say number of orders have increased from 2017 to 2018

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

```sql
select
count(order_id) as `number_of_orders`,
EXTRACT(MONTH FROM  order_purchase_timestamp) as `month`,
EXTRACT(YEAR FROM order_purchase_timestamp) as `year`
FROM
target.orders
group by
year, month
order by month, year
```

Row	number_of_orders	month	year
1	800	1	2017
2	7269	1	2018
3	1780	2	2017
4	6728	2	2018
5	2682	3	2017
6	7211	3	2018
7	2404	4	2017
8	6939	4	2018
9	3700	5	2017
10	6873	5	2018

Insight: It is pretty hard to check for any seasonality with the given data. In the months between May and Sept, there is a spike in orders.

C. 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

```sql
select
count(order_id) as `number_of_orders`,
(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'
  ELSE "Night"
  END) as `hour`
FROM
target.orders
GROUP BY
(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'
  ELSE "Night"
  END)
order by `number_of_orders`
```

Row	number_of_orders	hour
1	5242	Dawn
2	27733	Mornings
3	28331	Night
4	38135	Afternoon

Insight: Brazilian customers place their orders maximum during Afternoon.

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

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

Month-on-Month Analysis for year 2016

```sql
select
state,
month,
year,
sum(no_of_orders) as `total_orders`
from
(select
 customer_state as `state`,
extract(month from order_delivered_customer_date) as `month`,
extract(year from order_delivered_customer_date) as `year`,
count(order_id) over (partition by extract(month from order_delivered_cust
order by c.customer_id) as `no_of_orders`,
from
`target.orders` as `o` join `target.customers` as `c`
on o.customer_id = c.customer_id
where extract(year from order_delivered_customer_date) = 2016 OR
extract(year from order_delivered_customer_date) = 2017 OR
extract(year from order_delivered_customer_date) = 2018) as `T`
group by
state, month, year
order by
month
```


Row	state	month	year	total_orders
1	MG	1	2018	26
2	SP	1	2018	89
3	SC	1	2018	8
4	RS	1	2018	12
5	RJ	1	2018	30
6	ES	1	2018	4
7	PE	1	2018	4
8	MA	1	2018	2
9	MG	1	2017	1
10	BA	1	2018	8

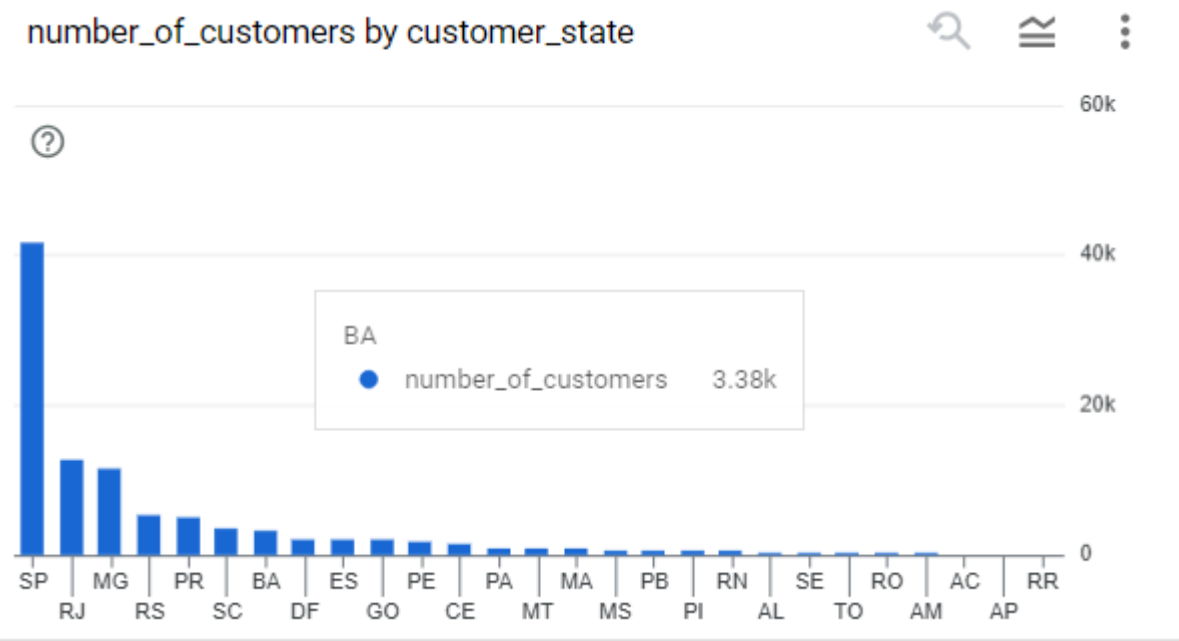
Insight: Again, month-on-month analyzing the trend in the number of orders placed over the past years is essential for understanding the business's growth trajectory. By examining the order volume month by month, we can identify whether there is a consistent upward trend indicating growth or not. Also, it will help us to check for seasonality in the orders placed In either increasing or decreasing order. This insight is crucial for strategic planning, resource allocation, and forecasting future business performance.

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

Untitled 3 RUN SAVE DOWNLOAD

```
1  SELECT
2  customer_state,
3  count(customer_id) as `number_of_customers`
4  from
5  target.customers
6  group by
7  customer_state
8  order by
9  number_of_customers DESC,
10 customer_state
```

Row	customer_state	number_of_custome
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



Insight: Understanding the distribution of customers across all states provides valuable insights into the geographic reach and market penetration of the business.

State SP has maximum number of customers.

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

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

```sql
select
round(((max(T.total_cost)-min (T.total_cost))/min(T.total_cost) * 100), 2
as `cost_of_pct_increase`
from
(
select
distinct
EXTRACT(year FROM o.order_delivered_customer_date) as `year`,
sum(payment_value) over( order by EXTRACT(year FROM o.order_delivered_cus
AS `total_cost`
FROM
`target.payments` as `p` JOIN
`target.orders`as `o`
on o.order_id = p.order_id
WHERE
EXTRACT(MONTH FROM o.order_delivered_customer_date) BETWEEN 1 AND 8
AND
EXTRACT(year FROM o.order_delivered_customer_date) = 2018
OR
EXTRACT(year FROM o.order_delivered_customer_date) = 2017) as `T`
```


Row	cost_of_pct_increase
1	135.94

Insight: Calculating the percentage increase in the cost of orders from year 2017 to 2018 provides valuable insights into the pricing dynamics and potential changes in customer spending behavior over time. By focusing on the months between January to August, we capture a significant portion of the annual business activity and can assess trends in order costs during this period.

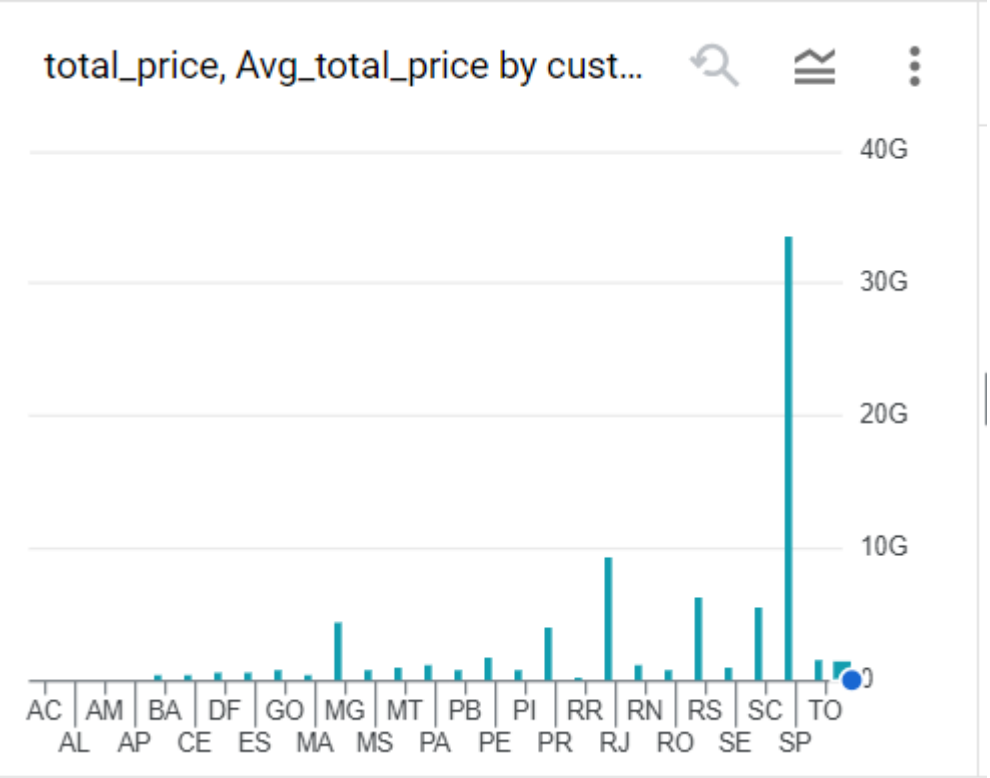
There is almost 136% increase the cost between 2017 and 2018.

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

```sql
select distinct
customer_state,
Round(sum(total_price) over(partition by customer_state), 2) as `total_pr
Round(AVG(total_price) over(partition by customer_state),2) as `Avg_total
from
(select distinct
c.customer_state,
sum(oi.price) over(partition by
extract(year from shipping_limit_date)
 order by c.customer_state) as `total_price`,
AVG(price) as `avg_price`,
extract(year from shipping_limit_date) as `year`
from
target.order_items as `oi` join target.orders as `o`
on oi.order_id=o.order_id join target.customers as `c`
on o.customer_id = c.customer_id
group by
customer_state, price, extract(year from shipping_limit_date),
shipping_limit_date) as `T`

order by
Avg_total_price, customer_state
```


Row	customer_state	total_price	Avg_total_price
1	AC	661317.22	8371.1
2	AL	14570110.94	45674.33
3	AM	7554077.24	56373.71
4	AP	3894678.63	61820.3
5	BA	406702935.02	298607.15
6	CE	295707274.01	400687.36
7	DF	541214725.89	546681.54
8	ES	652132763.13	675785.25
9	GO	780353445.17	806150.25
10	MA	421060347.4	861064.11



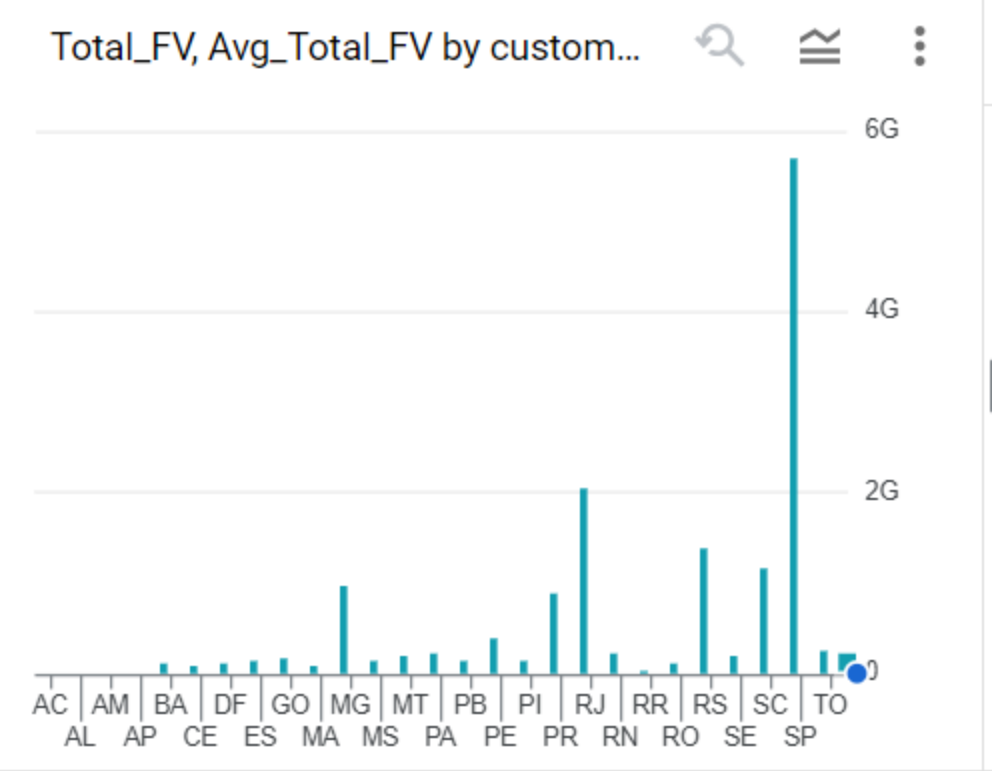
Insight: Calculating the total and average value of order prices for each state provides valuable insights into the revenue generated from different regions and the spending behavior of customers in each state. By analyzing the total order price, businesses can identify states with the highest and lowest revenue contributions, helps understand the purchasing power and preferences of customers in each state.

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

```sql
select distinct
customer_state,
Round(sum(total_freight_value) over(partition by customer_state), 2) as `
Round(AVG(total_freight_value) over(partition by customer_state),2) as `A
from
(select distinct
c.customer_state,
sum(oi.freight_value) over(partition by
extract(year from shipping_limit_date)
 order by c.customer_state) as `total_freight_value`,
AVG(freight_value) as `avg_freight_value`,
extract(year from shipping_limit_date) as `year`
from
target.order_items as `oi` join target.orders as `o`
on oi.order_id=o.order_id join target.customers as `c`
on o.customer_id = c.customer_id
group by
customer_state, freight_value, extract(year from shipping_limit_date),
shipping_limit_date) as `T`

order by
Avg_Total_FV, customer_state
```


Row	customer_state	Total_FV	Avg_Total_FV
1	AC	136747.98	1823.31
2	AL	3010219.21	9149.6
3	AM	1413953.11	11685.56
4	AP	828373.03	12943.33
5	BA	97532085.7	58472.47
6	CE	70385810.16	80074.87
7	DF	119476560.63	104804.0
8	ES	137211678.11	127283.56
9	GO	169141666.93	151019.35
10	MA	88828232.1	164496.73



Insight: Understanding the total and average order freight costs per state is crucial for logistics and operational planning. By analyzing these metrics, businesses can identify states with higher shipping costs and potential inefficiencies in their logistics operations.

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

A. 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.

```
1  select
2  order_id,
3  date_diff(order_delivered_customer_date, order_purchase_timestamp, day)
4  as `time_to-deliver`,
5  date_diff(order_estimated_delivery_date, order_delivered_customer_date, d
6  as `diff-estimated_delivery`
7  from
8  target.orders
9
```

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

Insight: Analyzing the delivery time and the difference between estimated and actual delivery dates provides valuable insights into the efficiency and reliability of the logistics operations. Understanding the factors contributing to delivery delays enables businesses to optimize their supply chain processes, address bottlenecks, and streamline operations to ensure timely order fulfillment.

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

Top 5 states with highest freight value

```sql
select distinct
customer_state,
sum(total_freight_value) over(partition by customer_state) as `Total_FV`,
AVG(total_freight_value) over(partition by customer_state) as `Avg_Total_
from
(select distinct
c.customer_state,
sum(oi.freight_value) over(partition by
extract(year from shipping_limit_date)
 order by c.customer_state) as `total_freight_value`,
AVG(freight_value) as `avg_freight_value`,
extract(year from shipping_limit_date) as `year`
from
target.order_items as `oi` join target.orders as `o`
on oi.order_id=o.order_id join target.customers as `c`
on o.customer_id = c.customer_id
group by
customer_state, freight_value, extract(year from shipping_limit_date),
shipping_limit_date) as `T`

order by
Avg_Total_FV DESC, customer_state
LIMIT 5
```


Row	customer_state	Total_FV	Avg_Total_FV
1	TO	238568673.39	1032764.819870…
2	SP	5696957156.97	1012432.407494…
3	SC	1153057103.409…	689215.2441183…
4	SE	186111640.7600…	679239.5648175…
5	RS	1394664573.84	642110.7614364…

Top 5 states with lowest freight value

```sql
select distinct
customer_state,
sum(total_freight_value) over(partition by customer_state) as `Total_FV`,
AVG(total_freight_value) over(partition by customer_state) as `Avg_Total_
from
(select distinct
c.customer_state,
sum(oi.freight_value) over(partition by
extract(year from shipping_limit_date)
 order by c.customer_state) as `total_freight_value`,
AVG(freight_value) as `avg_freight_value`,
extract(year from shipping_limit_date) as `year`
from
target.order_items as `oi` join target.orders as `o`
on oi.order_id=o.order_id join target.customers as `c`
on o.customer_id = c.customer_id
group by
customer_state, freight_value, extract(year from shipping_limit_date),
shipping_limit_date) as `T`

order by
Avg_Total_FV, customer_state
LIMIT 5
```


Row	customer_state	Total_FV	Avg_Total_FV
1	AC	136747.9799999...	1823.3064
2	AL	3010219.21	9149.602462006...
3	AM	1413953.109999...	11685.56289256...
4	AP	828373.0299999...	12943.32859374...
5	BA	97532085.7	58472.473441247

customer_state

Insight: Identifying the top 5 states with the highest and lowest average freight values provides valuable insights into regional disparities in shipping costs. Businesses can use this information to optimize their logistics strategies, negotiate better freight rates, and allocate resources effectively to minimize shipping expenses.

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

Top 5 states with highest average delivery time

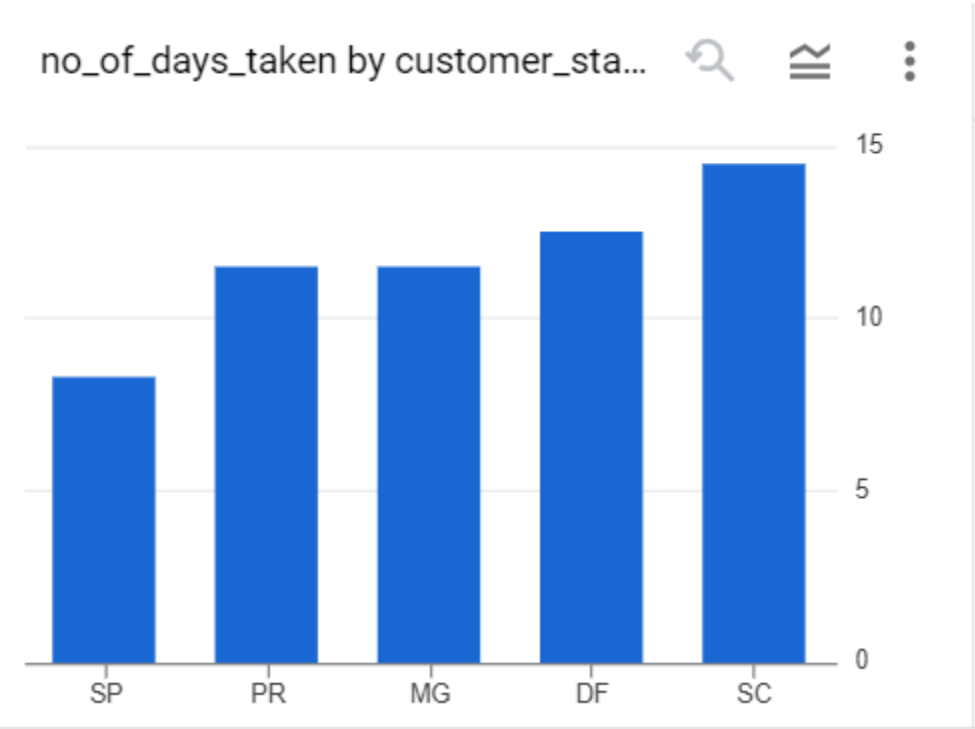
```sql
select
customer_state,
ROUND(AVG(date_diff(order_delivered_customer_date, order_purchase_timestar
as `no_of_days_taken`,
from
target.orders as `o` join target.customers as `c`
on o.customer_id = c.customer_id
group by
customer_state
order by
no_of_days_taken DESC, customer_state
LIMIT 5
```

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

Top 5 states with highest average delivery time

```sql
select
customer_state,
ROUND(AVG(date_diff(order_delivered_customer_date, order_purchase_timesta
as `no_of_days_taken`,
from
target.orders as `o` join target.customers as `c`
on o.customer_id = c.customer_id
group by
customer_state
order by
no_of_days_taken, customer_state
LIMIT 5
```

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



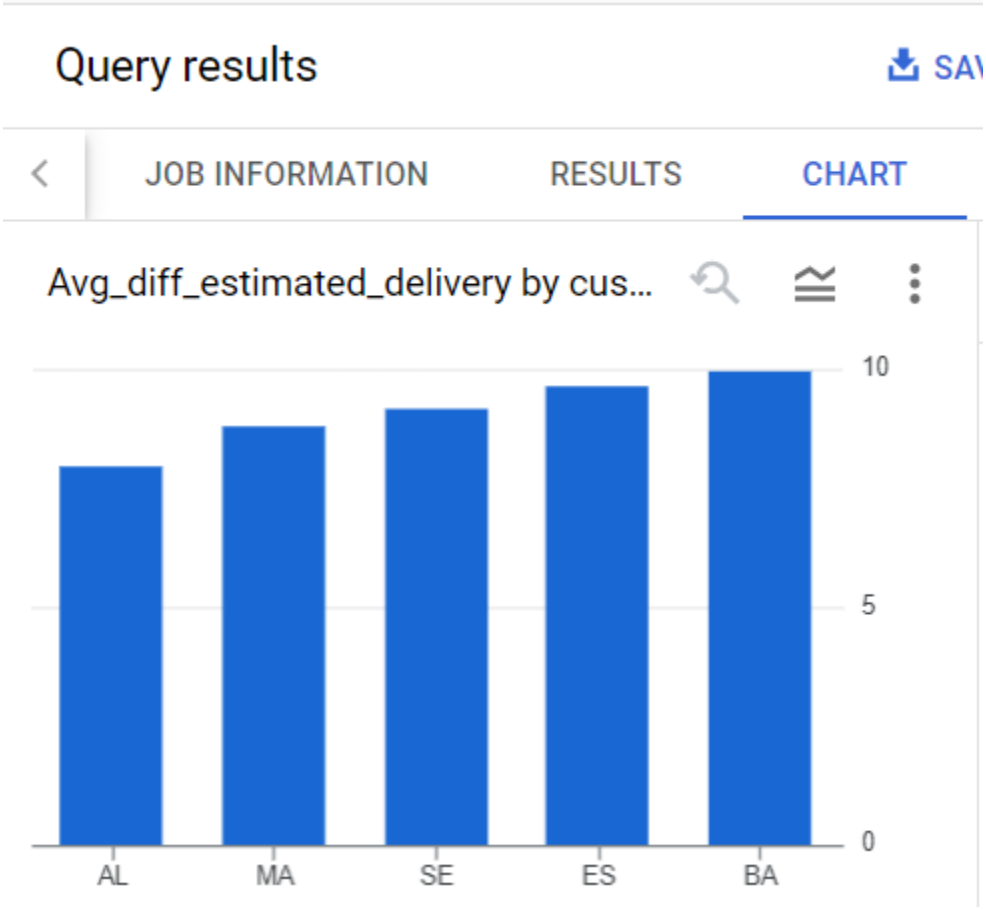
Insight: Identifying the top 5 states with the highest and lowest average delivery times provides insights into the efficiency of logistics operations and the overall customer experience in different regions.

D. Find out the top 5 states where the order delivery is really fast as compared to the estimated date of delivery.

```sql
select
customer_state,
round(AVG(date_diff(order_estimated_delivery_date, order_delivered_custome
2) as `Avg_diff_estimated_delivery`
from
target.orders as `o` join target.customers as `c`
on o.customer_id = c.customer_id
group by
customer_state
order by
Avg_diff_estimated_delivery, customer_state
LIMIT 5
```

Press Alt+F1

Row	customer_state	Avg_diff_estimated_
1	AL	7.95
2	MA	8.77
3	SE	9.17
4	ES	9.62
5	BA	9.93



Insight: Identifying the top 5 states where order delivery is faster than the estimated date of delivery provides valuable insights into the efficiency of logistics operations and the ability to meet or exceed customer expectations. Additionally, understanding the factors contributing to faster delivery times in specific states enables businesses to implement best practices and improve overall delivery performance across the board.

6. Analysis based on the payments:

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

```sql
select distinct
extract(month from order_delivered_customer_date) as `month`,
extract(year from order_delivered_customer_date) as `payment_year`,
p.payment_type,
count(p.order_id) over(partition by payment_type order by
extract(month from order_delivered_customer_date), extract(year from
order_delivered_customer_date)) as `no_of_orders`,
from
target.payments as `p` join target.orders as `o`
on p.order_id = o.order_id
order by
month
```

Row	month	payment_year	payment_type	no_of_ord
1	*null*	*null*	UPI	
2	*null*	*null*	credit_card	
3	*null*	*null*	debit_card	
4	*null*	*null*	not_defined	
5	*null*	*null*	voucher	
6	1	2017	UPI	
7	1	2018	UPI	
8	1	2017	credit_card	
9	1	2018	credit_card	
10	1	2017	debit_card	

Insight: Analyzing the month-on-month number of orders placed using different payment types provides insights into customer payment preferences and trends over time. It allows businesses to understand which payment methods are most popular among customers and how their usage may vary seasonally or in response to external factors such as promotions or economic conditions. This information can help businesses tailor their payment processing systems, optimize payment options, and develop targeted marketing strategies to better serve customer needs and preferences.

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

```sql
select distinct
count(payment_sequential) as `no_of_orders`
from
target.payments
where payment_sequential >=1
```

Row	no_of_orders
1	103886

Insights: Analyzing the number of orders placed based on the number of payment installments that have been paid provides insights into customer payment behaviors and preferences. It allows businesses to understand how customers choose to pay for their orders, whether they prefer to pay in full upfront or spread payments over multiple installments.

The payment_sequential column gives the sequences of “installments paid.” So count on it will enable to get number of order placed on the basis of installments that have been paid.

Overall recommendations

Optimize Logistical Operations: Implement best practices observed in the states with faster delivery time and efficient operations across the entire logistics network to improve overall delivery performance and enhance customer satisfaction.

Tailor Payment Options: Payment preferences and trends should be used to optimize payment processing systems and offer flexible payment options that align with customer preferences. Consider promoting payment methods that are popular among customers and streamline the checkout process to enhance the overall shopping experience.

Targeted Marketing Campaigns: Using geographic and historical time line insights one should do tailor marketing campaigns and promotions to specific regions and seasons. Customer distribution data should be used to identify high-value regions and implement targeted marketing strategies to drive customer engagement and sales.

Enhance Customer Service: Delivery times and customer feedback to identify areas for improvement in customer service and order fulfillment processes. Need to focus on reducing delivery times, addressing delivery delays, and enhancing communication with customers to provide a seamless shopping experience and improve overall customer satisfaction.

Data-Driven Decision Making: Continuously monitor key metrics such as order volume, delivery times, payment preferences, and customer feedback to make informed business decisions. Use data analytics to identify trends, opportunities, and areas for improvement, and adjust strategies accordingly to stay competitive and drive business growth.

By implementing these recommendations, businesses can optimize operations, improve customer satisfaction, and drive growth in the highly competitive retail landscape.