Problem Statement:

Assuming you are a data analyst/ scientist at Target, you have been assigned the task of analyzing the given dataset to extract valuable insights and provide actionable recommendations.

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

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

Ans:

Query:

```
SELECT
    column_name,
    data_type
From hybrid-life-444314-u7.Target_Company_Project.INFORMATION_SCHEMA.COLUMNS
WHERE table_name = 'customers';
```

Execution & Result:

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:

Reference to the details exploratory for the customers table found to be more String Type values and the integer is once.

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

Ans:

Query:

```
select
min(order_purchase_timestamp) as first_order_Placed,
max(order_purchase_timestamp) as last_order_Placed,
from `Target_Company_Project.orders`;
```

Execution & Result:

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

Insights:

By analyzing the orders range between 2016 and 2018 and found that fist order were placed in given time period is on 04/09/2016 and last order placed on 17-10-2018.

Problem Statement:

Assuming you are a data analyst/ scientist at Target, you have been assigned the task of analyzing the given dataset to extract valuable insights and provide actionable recommendations.

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

Ans:

Query:

```
select
count(distinct customer_city) as customer_city_count,
count(distinct customer_state) as customer_State_count
from `Target_Company_Project.orders` o
JOIN `Target_Company_Project.customers` c
On o.customer_id = c.customer_id;
```

Execution & Result:

Row	customer_city_count	customer_State_count
1	4119	27

Insights:

By analyzing customers and their states during the given period is that total no of states is 27 and total no of city of their customers is 4119.

II. In-depth Exploration:

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

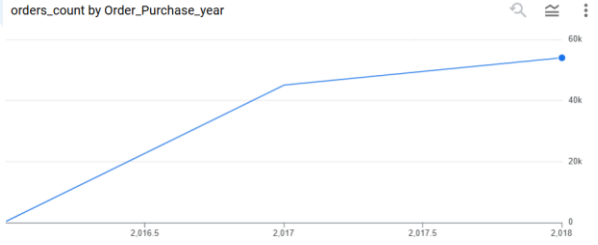
Query

```
select
extract(year from order_purchase_timestamp) as Order_Purchase_year,
count(*) as orders_count,
DENSE_RANK() OVER (ORDER BY COUNT(*) DESC) AS Rank_By_Year
from `Target_Company_Project.orders`
group by 1
order by 1;
```

Execution & Result:

Row	Order_Purchase_year	orders_count	Rank_By_Year
1	2016	329	3
2	2017	45101	2
3	2018	54011	1

Insights:



By Analyzing the growing trend in no.of orders placed during the time period 2016 to 2018 is **high** and the **Margin** of growth is too, To simply the analyzing trend given is the graph to the left and also given Rank by the year in Execution.

Problem Statement:

Assuming you are a data analyst/ scientist at Target, you have been assigned the task of analyzing the given dataset to extract valuable insights and provide actionable recommendations.

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

Query

```
select
extract(year from order_purchase_timestamp) as year,
extract(month from order_purchase_timestamp) as month,
count(*) as orders_count
from `Target_Company_Project.orders`
group by 1,2
order by 1,2
Limit 10;
```

Execution & Result:

Row	year	month	orders_count
1	2016	9	4
2	2016	10	324
3	2016	12	1
4	2017	1	800
5	2017	2	1780
6	2017	3	2682
7	2017	4	2404
8	2017	5	3700
9	2017	6	3245
10	2017	7	4026

Note: Total output 25 rows, Limited by Just 10 for comfort view.

Insights:

->March 2018 (Orders: 7,211) had the highest order count, indicating a possible seasonal or promotional activity during that period.

-> April 2018 and May 2018 also had high order counts (6,939 and 6,873, respectively), suggesting sustained customer engagement during the spring season.

->Starting from June 2018 (Orders: 6,167), there is a visible decline in the number of orders month-over-month, which continues into July (6,292) and August (6,512).

Recommendation:

- Identify factors contributing to the **sharp decline after August**.
- Examine if **seasonal trends, competition, or operational issues** played a role.
- Consider offering promotions or incentives to regain market share

Problem Statement:

Assuming you are a data analyst/ scientist at Target, you have been assigned the task of analyzing the given dataset to extract valuable insights and provide actionable recommendations.

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

Query

```
select
 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_order_placed,
 count(*) as No_of_orders_Placed,
 DENSE_RANK() OVER (ORDER BY COUNT(*) DESC) AS Rank_by_time_of_Order_Placed
from `Target_Company_Project.orders`
group by 1
order by 2 desc;
```

Execution and Result:

Row	Time_of_day_order_placed	No_of_orders_Placed	Rank_by_time_of_Order_Placed
1	Afternoon	38135	1
2	Night	28331	2
3	Mornings	27733	3
4	Dawn	5242	4

Insights:

By Analyzing the Brazilian customers mostly place their orders in **Afternoon** of the day with order in numbers (38,135), The above query execution results shows during afternoon is the highest order placing time whereas Dawn time is the lowest ordering time with orders in numbers are(5242).

Recommendations:

Warehouse & Delivery Optimization:

More orders in the afternoon → Ensure warehouses are well-staffed & delivery logistics are optimized during peak order hours.

Offer same-day or express delivery for afternoon and night orders.

Use AI-driven personalization to suggest products based on when customers typically shop.

Offer limited-time "Night Owl" or "Early Bird" discounts to drive engagement during off-peak hours.

Problem Statement:

Assuming you are a data analyst/ scientist at Target, you have been assigned the task of analyzing the given dataset to extract valuable insights and provide actionable recommendations.

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

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

Query

```
select
count(*) as no_of_orders,
extract(month from order_purchase_timestamp) as orders_month_In,
customer_state
from `Target_Company_Project.orders` o
JOIN `Target_Company_Project.customers` c
USING(customer_id)
group by 3,2
order by 1 desc
Limit 10;
```

Execution and Result:

Row	no_of_orders	orders_month_In	customer_state
1	4982	8	SP
2	4632	5	SP
3	4381	7	SP
4	4104	6	SP
5	4047	3	SP
6	3967	4	SP
7	3357	2	SP
8	3351	1	SP
9	3012	11	SP
10	2357	12	SP

Note: Total output 322 rows, Limited by Just 10 for comfort view.

Insights:

RR has the highest single-month order count with **4 orders**, indicating that April might be a strong season for this state.

Focus on RR:

- RR is a top-performing state, but its contributions vary across months. Strategies can be deployed to boost its performance in low-activity months like January and November.

Problem Statement:

Assuming you are a data analyst/ scientist at Target, you have been assigned the task of analyzing the given dataset to extract valuable insights and provide actionable recommendations.

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

Query:

```
select
customer_state,
extract(year from order_purchase_timestamp) as order_year_In,
extract(month from order_purchase_timestamp) as orders_month_In,
count(*) as no_of_orders
from `Target_Company_Project.orders` o
JOIN `Target_Company_Project.customers` c
ON o.customer_id = c.customer_id
group by 1,2,3
order by 4 desc
Limit 10;
```

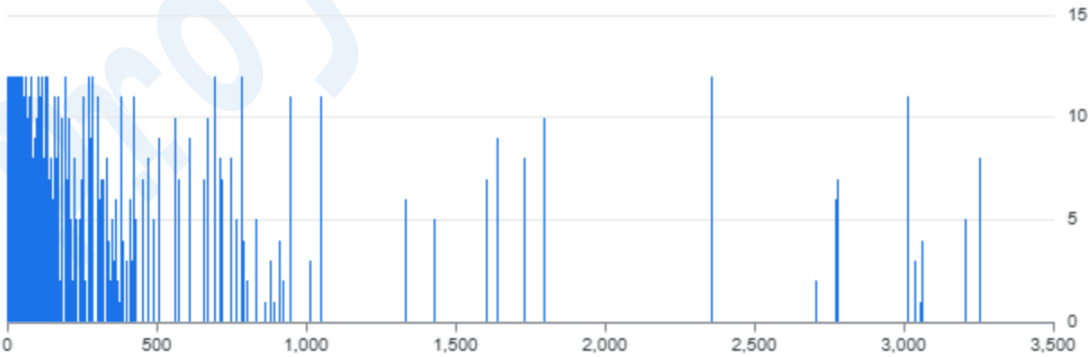
Execution & Result:

Row	customer_state	order_year_In	orders_month_In	no_of_orders
1	SP	2018	8	3253
2	SP	2018	5	3207
3	SP	2018	4	3059
4	SP	2018	1	3052
5	SP	2018	3	3037
6	SP	2017	11	3012
7	SP	2018	7	2777
8	SP	2018	6	2773
9	SP	2018	2	2703
10	SP	2017	12	2357

Note: Total output 565 rows, Limited by Just 10 for comfort view.

Insights:

orders_month_In by no_of_orders



Focus on Regional Growth:
States like **RR** with consistent activity could benefit from focused campaigns to strengthen engagement.
States like **TO**, **PB**, and **PI**, which show limited activity, could be targeted for customer acquisition.

Recommendations;

Ensure higher stock availability during peak order months.

Consider warehouse locations near high-order regions (e.g., SP, RJ, MG) to reduce delivery time and costs.

Problem Statement:

Assuming you are a data analyst/ scientist at Target, you have been assigned the task of analyzing the given dataset to extract valuable insights and provide actionable recommendations.

IV. 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).*

You can use the "payment_value" column in the payments table to get the cost of orders.

Query:

```
with percentage AS
(
select
*
from `Target_Company_Project.orders` o
JOIN `Target_Company_Project.payments` p
On o.order_id = p.order_id
where extract(year from order_purchase_timestamp) BETWEEN 2017 AND 2018
AND extract(month from order_purchase_timestamp) BETWEEN 1 AND 8
),
percentage1 AS
(
 select
 extract(year from order_purchase_timestamp) as year,
 ROUND(sum(payment_value),2) as cost
 from percentage
 group by year
 order by 1
),
percentage2 AS
(
select
 *,LEAD(cost) OVER(order by year) as nxt_year_cost
 from percentage1
)
select *,round((nxt_year_cost - cost) / cost * 100,2) as percentage_increased
from percentage2;
```

Execution & Result:

Row	year	cost	next_year_cost	percentage_increased
1	2017	3669022.12	8694733.84	136.98
2	2018	8694733.84	*null*	*null*

Insights:

By Analyzing the data it is found that to be in 2018 **136.98%** is increased compared to 2017.

Problem Statement:

Assuming you are a data analyst/ scientist at Target, you have been assigned the task of analyzing the given dataset to extract valuable insights and provide actionable recommendations.

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

Query:

```
select
 customer_state,
 ROUND(sum(price),2) as total_price_by_states,
 ROUND(avg(price),2) as average_price_by_states,
from `Target_Company_Project.orders` o
JOIN `Target_Company_Project.order_items` oi
ON o.order_id = oi.order_id
JOIN `Target_Company_Project.customers` c
ON o.customer_id = c.customer_id
group by customer_state
order by 2
Limit 10;
```

Execution and Result:

Row	customer_state	total_price_by_states	average_price_by_states
1	RR	7829.43	150.57
2	AP	13474.3	164.32
3	AC	15982.95	173.73
4	AM	22356.84	135.5
5	RO	46140.64	165.97
6	TO	49621.74	157.53
7	SE	58920.85	153.04
8	AL	80314.81	180.89
9	RN	83034.98	156.97
10	PI	86914.08	160.36

Note: Total output 27 rows, Limited by Just 10 for comfort view.

Insights:

By Analyzing RR is the lowest in price by states and AP and AC is followed by them, Whereas seems that average price is equally share ratio except by AM state(135.5).

Recommendations.

Prioritize marketing efforts and pricing strategies in states with high total sales but low average prices (e.g., SP, MG, RS, SC). These states have strong sales volume but lower per-unit pricing, indicating potential opportunities for price optimization or premium product offerings.

Problem Statement:

Assuming you are a data analyst/ scientist at Target, you have been assigned the task of analyzing the given dataset to extract valuable insights and provide actionable recommendations.

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

Query:

```
select
 customer_state,
 ROUND(sum(freight_value),2) as total_freight_price_by_states,
 ROUND(avg(freight_value),2) as average_freight_price_by_states,
from `Target_Company_Project.orders`o
JOIN `Target_Company_Project.order_items`oi
ON o.order_id = oi.order_id
JOIN `Target_Company_Project.customers`c
ON o.customer_id = c.customer_id
group by customer_state
order by 2
Limit 10;
```

Execution and Result:

Row	customer_state	total_freight_price_by_states	average_freight_price_by_states
1	RR	2235.19	42.98
2	AP	2788.5	34.01
3	AC	3686.75	40.07
4	AM	5478.89	33.21
5	RO	11417.38	41.07
6	TO	11732.68	37.25
7	SE	14111.47	36.65
8	AL	15914.59	35.84
9	RN	18860.1	35.65
10	MS	19144.03	23.37

Note: Total output 27 rows, Limited by Just 10 for comfort view.

Insights:

SP State leads significantly in total freight price with **718,723.07**, making up the majority share. However, it has the **lowest average freight price** (**15.15**), indicating a large number of smaller shipments.

PE State has the **highest average freight price** (**32.92**), indicating higher shipping costs,

Recommendations:

Based on the data provided, a recommendation would be to focus on optimizing freight pricing for the states with the highest total freight prices, particularly SP, RJ, and MG, as these states have the largest total freight amounts. While their average freight prices are lower compared to others, these states represent a significant portion of the total freight volume.

Problem Statement:

Assuming you are a data analyst/ scientist at Target, you have been assigned the task of analyzing the given dataset to extract valuable insights and provide actionable recommendations.

V.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." 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

Query:

```
select
 order_id,
 TIMESTAMP_DIFF(order_delivered_customer_date,order_purchase_timestamp,DAY) as
Committed_time_to_deliver,
 TIMESTAMP_DIFF(order_delivered_customer_date,order_estimated_delivery_date,DAY)
as diff_estimated_delivery
from `Target_Company_Project.orders`
where order_status = 'delivered'
order by Committed_time_to_deliver desc
Limit 10;
```

Execution and Result:

Row	order_id	Committed_time_to_deliver	diff_estimated_delivery
1	ca07593549f1816d26a572e06...	209	181
2	1b3190b2dfa9d789e1f14c05b...	208	188
3	440d0d17af552815d15a9e41a...	195	165
4	0f4519c5f1c541ddec9f21b3bd...	194	161
5	2fb597c2f772eca01b1f5c561b...	194	155
6	285ab9426d6982034523a855f...	194	166
7	47b40429ed8cce3aee9199792...	191	175
8	2fe324febf907e3ea3f2aa9650...	189	167
9	2d7561026d542c8dbd8f0daea...	188	159
10	c27815f7e3dd0b926b5855262...	187	162

Total output 96,478 rows, Limited by Just 10 for comfort view.
Committed_time_to_delivery is order in desc

Insights:

By Analyzing the order delivery committed date vs estimated delivery date the below is the Table of numbers in facts

Delivery	Count of Delivery Remarks	%
Early	87182	90%
Late	6534	7%
On Time	2754	3%
Grand Total	**96470**	100%

Recommendations:

Logistics Team needs to improve to cut-off delivery time deficit of 7% across all orders.

Problem Statement:

Assuming you are a data analyst/ scientist at Target, you have been assigned the task of analyzing the given dataset to extract valuable insights and provide actionable recommendations.

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

Highest Average Freight Value;

Query 1:

```
select
 customer_state,
 ROUND(avg(freight_value),2) as High_avg_freight_price,
from `Target_Company_Project.orders` o
JOIN `Target_Company_Project.order_items` oi
ON o.order_id = oi.order_id
JOIN `Target_Company_Project.customers` c
ON o.customer_id = c.customer_id
group by customer_state
order by 2 desc
LIMIT 5
```

Execution and Result:

Row	customer_state	High_avg_freight_price
1	RR	42.98
2	PB	42.72
3	RO	41.07
4	AC	40.07
5	PI	39.15

Query 2: Lowest Average Freight Value.

```
select
 customer_state,
 ROUND(avg(freight_value),2) as Low_average_freight_price,
from `Target_Company_Project.orders` o
JOIN `Target_Company_Project.order_items` oi
ON o.order_id = oi.order_id
JOIN `Target_Company_Project.customers` c
ON o.customer_id = c.customer_id
group by customer_state
order by 2
LIMIT 5
```

Execution and Result:

Row	customer_state	Low_avgerage_freight_price
1	SP	15.15
2	PR	20.53
3	MG	20.63
4	RJ	20.96
5	DF	21.04

Problem Statement:

Assuming you are a data analyst/ scientist at Target, you have been assigned the task of analyzing the given dataset to extract valuable insights and provide actionable recommendations.

V.B Combined:

Query 3:

```
(
select
 customer_state,
 ROUND(avg(freight_value),2) as High_to_avg_freight_price,
from `Target_Company_Project.orders` o
INNER JOIN `Target_Company_Project.order_items` oi
ON o.order_id = oi.order_id
INNER JOIN `Target_Company_Project.customers` c
ON o.customer_id = c.customer_id
group by customer_state
order by 2 desc
LIMIT 5
)
UNION ALL
(
select
 customer_state,
 ROUND(avg(freight_value),2) as High_to_Low_avg_freight_price,
from `Target_Company_Project.orders` o
INNER JOIN `Target_Company_Project.order_items` oi
ON o.order_id = oi.order_id
INNER JOIN `Target_Company_Project.customers` c
ON o.customer_id = c.customer_id
group by customer_state
order by 2
LIMIT 5
);
```

Row	customer_state	High_to_Low_avg_freight_price
1	RR	42.98
2	PB	42.72
3	RO	41.07
4	AC	40.07
5	PI	39.15
6	SP	15.15
7	PR	20.53
8	MG	20.63
9	RJ	20.96
10	DF	21.04

Problem Statement:

Assuming you are a data analyst/ scientist at Target, you have been assigned the task of analyzing the given dataset to extract valuable insights and provide actionable recommendations.

Insights:

Query1: Found Top Five States for which Highest Average Freight Value and the Lowest Average Freight Value across the states,

Recommendations:

Focus on optimizing operational costs in states with high average freight prices, such as RR, PB, and RO, to enhance profitability while maintaining service quality.

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

Query:

```
With CTE_AVG_DEV_TIME as
( select C.customer_state as State, round(avg(date_diff(O.order_delivered_customer_date, O.order_purchase_timestamp, DAY)),2) as Avg_Del_Time
from `hybrid-life-444314-u7.Target_Company_Project.orders` O
join `hybrid-life-444314-u7.Target_Company_Project.customers` C on O.customer_id = C.customer_id
where O.order_delivered_customer_date is not null
group by 1 ),
CTE_STATE_RNK as
(
  select State, Avg_Del_Time ,
  dense_rank() over(order by Avg_Del_Time Desc) as H_RANK,
  dense_rank() over(order by Avg_Del_Time ) as L_RANK
  from CTE_AVG_DEV_TIME
)
select H.State as High_Delivery_Time,
L.State as Low_Delivery_State
from CTE_STATE_RNK H join CTE_STATE_RNK L on H.H_Rank = L.L_Rank
where H.H_Rank <=5;
```

Execution and Result:

Row	High_Delivery_Time	Low_Delivery_State
1	AL	DF
2	AP	PR
3	AM	MG
4	RR	SP
5	PA	SC

Insights:

States with high delivery times (e.g., AL, AP, AM) are geographically distant or face logistical challenges compared to low-delivery-time states like DF and SP.

Recommendations:

Invest in optimizing supply chain infrastructure and last-mile delivery strategies for high-delivery-time states to reduce delays and improve customer satisfaction.

Problem Statement:

Assuming you are a data analyst/ scientist at Target, you have been assigned the task of analyzing the given dataset to extract valuable insights and provide actionable recommendations.

V D. 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
customer_state AS state,
ROUND(SUM(TIMESTAMP_DIFF(order_delivered_customer_date, order_purchase_timestamp,
DAY))/COUNT(ORDER_ID), 2) AS average_time_for_delivery,
ROUND(SUM(TIMESTAMP_DIFF(order_estimated_delivery_date, order_purchase_timestamp,
DAY))/COUNT(ORDER_ID), 2) AS average_estimated_delivery_time,
FROM `Target_Company_Project.orders` o
JOIN `Target_Company_Project.customers` c
ON o.customer_id=c.customer_id
WHERE order_status='delivered'
GROUP BY customer_state
ORDER BY (average_time_for_delivery-average_estimated_delivery_time)
Limit 5;
```

Execution and Result:

Row	state	average_time_for_delivery	average_estimated_delivery_time
1	AC	20.64	40.73
2	RO	18.91	38.39
3	AP	26.73	45.87
4	AM	25.99	44.92
5	RR	28.98	45.63

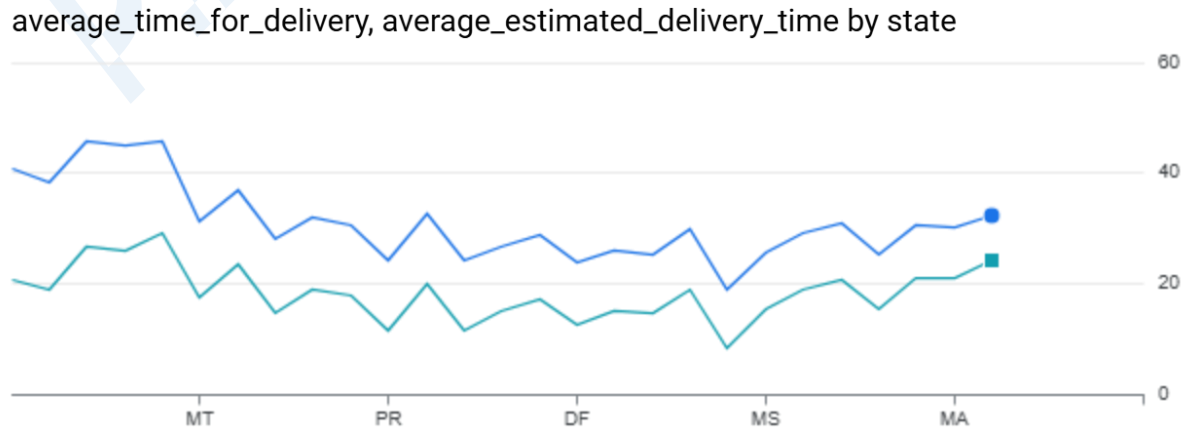
Insights:

States like SP, RJ, and PR have significantly lower actual and estimated delivery times, indicating efficient logistics, while states like RR and AP show higher delivery times, suggesting logistical challenges.

Recommendations:

Prioritize addressing logistical inefficiencies and infrastructure improvements in high-delivery-time states like RR and AP to align closer to the efficiency levels seen in SP and RJ.

Graph:



Problem Statement:

Assuming you are a data analyst/ scientist at Target, you have been assigned the task of analyzing the given dataset to extract valuable insights and provide actionable recommendations.

VI.Analysis based on the payments:

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

Query:

```
select
 payment_type,
 extract(year from order_purchase_timestamp) as year,
 extract(month from order_purchase_timestamp) as month,
 count(o.order_id) as number_of_orders_placed
from `Target_Company_Project.payments` p
JOIN `Target_Company_Project.orders` o
ON p.order_id = o.order_id
group by payment_type,year,month
order by 2,3
Limit 10;
```

Execution and Result:

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

Total output 90 rows, Limited by Just 10 for comfort view.

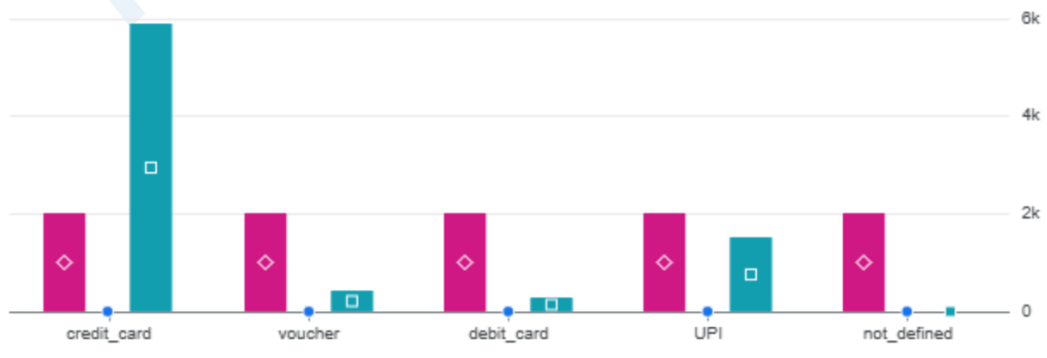
Insights:

Credit cards dominate as the most preferred payment type, with a consistent and significant increase in orders over time, while other methods like debit cards and vouchers remain less utilized.

Recommendations:

Focus on incentivizing underutilized payment methods like UPI and vouchers through promotions or discounts to diversify payment preferences and enhance customer convenience.

year, month, number_of_orders_placed by payment_type



Problem Statement:

Assuming you are a data analyst/ scientist at Target, you have been assigned the task of analyzing the given dataset to extract valuable insights and provide actionable recommendations.

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

Query:

```
select
 count(distinct order_id) as number_of_orders,
 payment_installments
from `Target_Company_Project.payments`
where payment_installments >= 1
group by payment_installments
LIMIT 10;
```

Execution and Result:

Row	number_of_orders	payment_installment
1	49060	1
2	12389	2
3	10443	3
4	7088	4
5	5234	5
6	3916	6
7	1623	7
8	4253	8
9	644	9
10	5315	10

Total output 23 rows, Limited by Just 10 for comfort view.

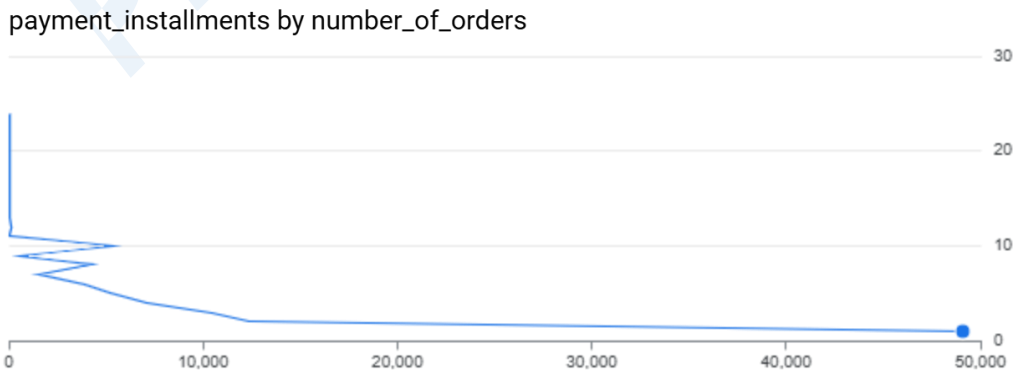
Insights:

The majority of orders are made with a single payment installment, while orders with higher installments decrease significantly, indicating a preference for simpler payment terms.

Recommendations:

Offer targeted promotions or interest-free installment plans to encourage customers to opt for higher installment options, potentially increasing the average order value.

Graph:



Problem Statement:

Assuming you are a data analyst/ scientist at Target, you have been assigned the task of analyzing the given dataset to extract valuable insights and provide actionable recommendations.

- Project Target Done by Mudassir Hasan N on 28-01-2025 @ 05:10 PM.
- Project Target was thoroughly analyze and given the best and comprehensive queries to extract the data. To optimize the space and time for the same.
- Project Target in each question and answers below to that insights and recommendations are given based on the available data.