TARGET CASE STUDY

- Target is a globally renowned brand and a prominent retailer in the United States. Target makes itself a preferred shopping destination by offering outstanding value, inspiration, innovation and an exceptional guest experience that no other retailer can deliver.
- This particular business case focuses on the operations of Target in Brazil and provides insightful information about 100,000 orders placed between 2016 and 2018. The dataset offers a comprehensive view of various dimensions including the order status, price, payment and freight performance, customer location, product attributes, and customer reviews.
- By analyzing this extensive dataset, it becomes possible to gain valuable insights into Target's operations in Brazil. The information can shed light on various aspects of the business, such as order processing, pricing strategies, payment and shipping efficiency, customer demographics, product characteristics, and customer satisfaction levels.

Dataset:
https://drive.google.com/drive/folders/1TGEc66YKbD443nslRi1bWgVd238gJCnb?usp=sharing

The data is available in 8 different csv files:

1. customers.csv
2. geolocation.csv
3. order_items.csv
4. payments.csv
5. reviews.csv
6. orders.csv
7. products.csv
8. sellers.csv

The column description for these csv files is given below.

The customers.csv contain following features:

Features	Description
customer_id	ID of the consumer who made the purchase
customer_unique_id	Unique ID of the consumer
customer_zip_code_prefix	Zip Code of consumer’s location

customer_city	Name of the City from where order is made
customer_state	State Code from where order is made (Eg. são paulo - SP)

The orders.csv contain following features:

Features	**Description**
order_id	A Unique ID of order made by the consumers
customer_id	ID of the consumer who made the purchase
order_status	Status of the order made i.e. delivered, shipped, etc.
order_purchase_timestamp	Timestamp of the purchase
order_delivered_carrier_date	Delivery date at which carrier made the delivery
order_delivered_customer_date	Date at which customer got the product
order_estimated_delivery_date	Estimated delivery date of the products

The order_items.csv contain following features:

Features	**Description**
order_id	A Unique ID of order made by the consumers
order_item_id	A Unique ID given to each item ordered in the order
product_id	A Unique ID given to each product available on the site
seller_id	Unique ID of the seller registered in Target
shipping_limit_date	The date before which the ordered product must be shipped
price	Actual price of the products ordered
freight_value	Price rate at which a product is delivered from one point to another

The payments.csv contain following features:

Features	Description
order_id	A Unique ID of order made by the consumers
payment_sequential	Sequences of the payments made in case of EMI
payment_type	Mode of payment used (Eg. Credit Card)
payment_installments	Number of installments in case of EMI purchase
payment_value	Total amount paid for the purchase order

The geolocations.csv contain following features:

Features	Description
geolocation_zip_code_prefix	First 5 digits of Zip Code
geolocation_lat	Latitude
geolocation_lng	Longitude
geolocation_city	City
geolocation_state	State

The sellers.csv contains following features:

Features	Description
seller_id	Unique ID of the seller registered
seller_zip_code_prefix	Zip Code of the seller’s location
seller_city	Name of the City of the seller
seller_state	State Code (Eg. são paulo - SP)

The reviews.csv contain following features:

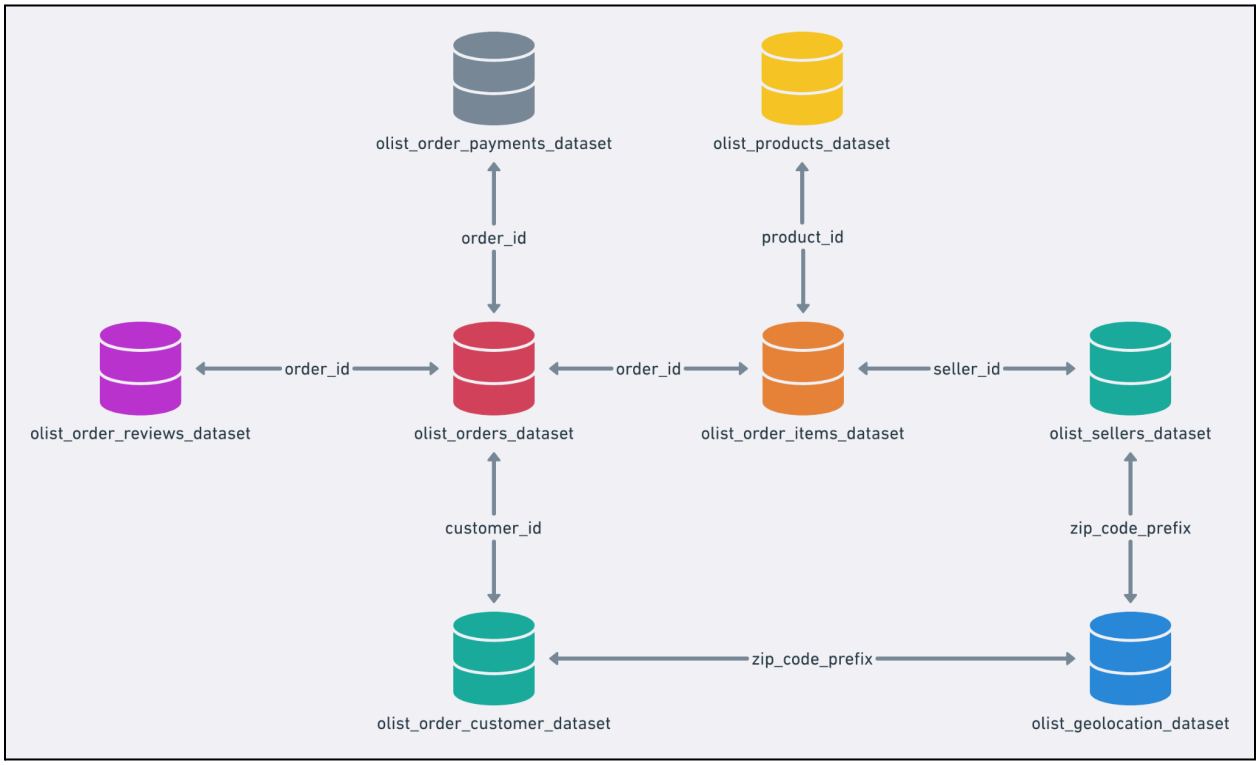
Features	Description
review_id	ID of the review given on the product ordered by the order id
order_id	A Unique ID of order made by the consumers
review_score	Review score given by the customer for each order on a scale of 1-5
review_comment_title	Title of the review
review_comment_message	Review comments posted by the consumer for each order
review_creation_date	Timestamp of the review when it is created
review_answer_timestamp	Timestamp of the review answered

The products.csv contain following features:

Features	Description
product_id	A Unique identifier for the proposed project.
product_category_name	Name of the product category
product_name_lenght	Length of the string which specifies the name given to the products ordered
product_description_lenght	Length of the description written for each product ordered on the site
product_photos_qty	Number of photos of each product ordered available on the shopping portal
product_weight_g	Weight of the products ordered in grams
product_length_cm	Length of the products ordered in centimeters
product_height_cm	Height of the products ordered in centimeters

product_width_cm	Width of the product ordered in centimeters

Dataset schema:



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.

Objective (SQL-based):

1. Extract data from the given dataset using SQL queries.
2. Explore table schemas and relationships to understand data structure.
3. Clean and transform data using SQL functions (e.g., filtering, joins, aggregations).
4. Analyze trends and patterns with SQL operations such as `GROUP BY`, `ORDER BY`, `COUNT`, `MAX`, `MIN`, and `AVG`.
5. Derive insights related to customer behavior, sales performance, and operations.
6. Provide actionable recommendations based on SQL query results to support business decisions at Target.

Data Analysis

1. 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.

Hint: We want you to display the data type of each column present in the “customers” table.

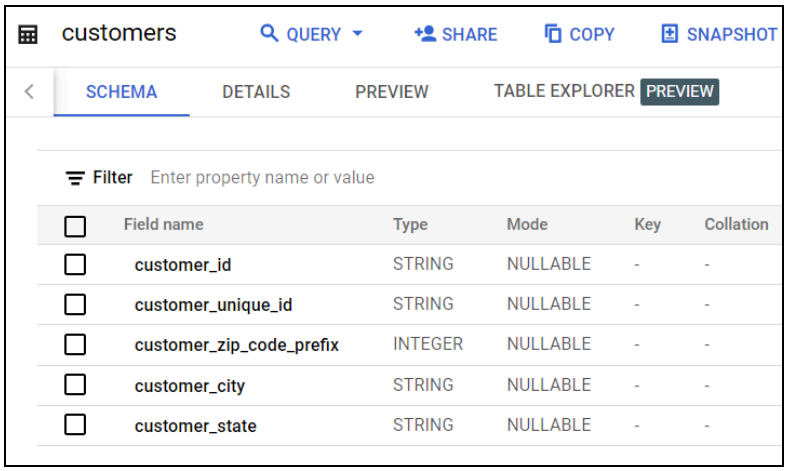
```
SELECT table_name, column_name, data_type
FROM Target_Dataset.INFORMATION_SCHEMA.COLUMNS
WHERE table_name = 'customers'
```

```
1
2  SELECT table_name, column_name, data_type
3  FROM Target_Dataset.INFORMATION_SCHEMA.COLUMNS
4  WHERE table_name = 'customers'
```

Query results

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS | EXECUTION

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



Inference & Insights:

To identify the structure of a given table, the **`information_schema`** was queried to retrieve the **data types of its columns.**

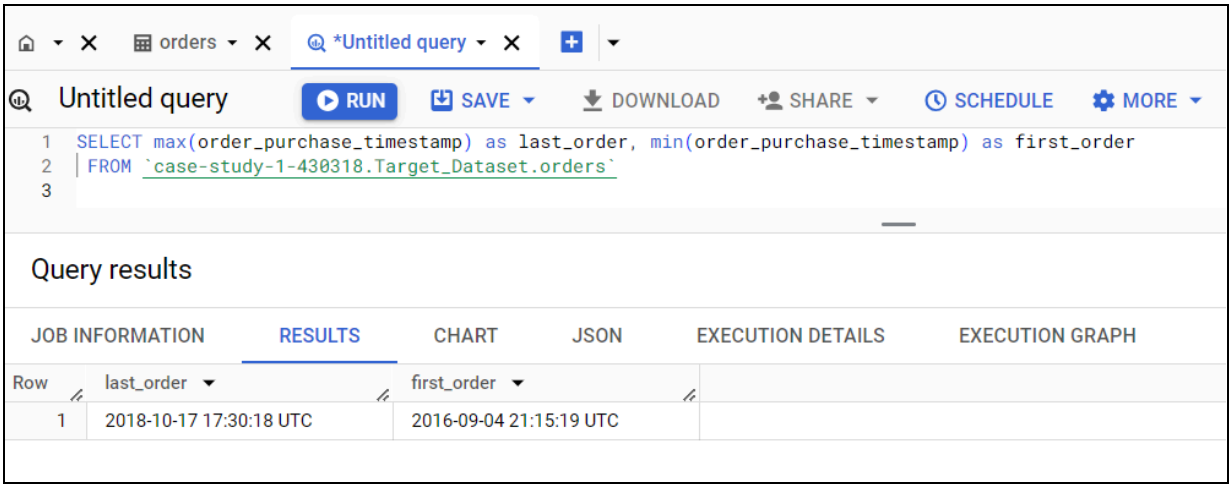
1. This approach ensures a clear understanding of the **table schema**, which is critical before performing transformations, aggregations, or applying constraints.
2. Knowing the **data type of each column** helps in validating data quality, optimizing queries, and avoiding type-related errors (e.g., performing arithmetic on string fields).
3. Schema exploration also assists in **data modeling** and ensures that downstream processes such as joins, aggregations, and visualizations are accurate and efficient.

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

Hint: We want you to get the date & time when the first and last orders in our dataset was placed.

```
SELECT max(order_purchase_timestamp) as last_order,
min(order_purchase_timestamp) as first_order
FROM `case-study-1-430318.Target_Dataset.orders` LIMIT 1000
```

PFB SS:



Inference & Insights:

By applying the MIN and MAX functions on the `order_purchase_timestamp` column from the orders table, it was observed that the first recorded order was placed in 2016 and the last order in 2018.

1. This indicates that the dataset spans a 3-year period (2016–2018), providing a sufficient timeline for analyzing purchasing trends, seasonality, and customer behavior.
2. The time range allows for year-over-year comparisons to evaluate growth, peak order months, and sales patterns.
3. Such insights can be leveraged to identify long-term trends, business cycles, and the effectiveness of marketing campaigns during the dataset period.

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

Hint: We want you to count the number of unique cities & states where orders were placed by the customers during the given time period.

```
SELECT
count(distinct c.customer_city) as city,
count(distinct c.customer_state) as state
FROM `case-study-1-430318.Target_Dataset.orders` o join
`case-study-1-430318.Target_Dataset.customers` c
on o.customer_id = c.customer_id
```

```sql

SELECT
count(distinct c.customer_city) as city,
count(distinct c.customer_state) as state
FROM `case-study-1-430318.Target_Dataset.orders` o
join `case-study-1-430318.Target_Dataset.customers` c
on o.customer_id = c.customer_id

```

Query results

JOB INFORMATION | RESULTS | CHART | JSON

Row	city	state
1	4119	27

Inference & Insights:

From the dataset analysis, it was observed that there are 4,119 unique cities spread across 27 states. This indicates wide geographical coverage and diversity in the dataset.

1. The high number of cities highlights granular-level representation, which can be leveraged for location-specific insights, decision-making, and regional trend analysis.
2. The relatively smaller number of states (27) compared to cities indicates that data is well-distributed across states, with each state contributing multiple city-level entries.
3. Such distribution enables deeper state-wise segmentation while still preserving city-level microanalysis, useful for understanding regional variations, infrastructure planning, or service distribution.

2. In-depth Exploration:

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

Hint: We want you to find out if no. of orders placed has increased gradually in each month, over the past years.

```sql
select year,
count(order_id) as order_count
from
(SELECT distinct extract(year from order_purchase_timestamp) as year, order_id
 FROM `case-study-1-430318.Target_Dataset.orders`
) X
group by year
order by year asc
```

```sql
select year,
count(order_id) as order_count
from
(SELECT distinct extract(year from order_purchase_timestamp) as year, order_id
 FROM `case-study-1-430318.Target_Dataset.orders`
) X
group by year
order by year asc
```

Query results

JOB INFORMATION | RESULTS | CHART | JSON | EXECUTION DETAILS

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

Another way:

```sql
select distinct year, month,
count(order_id) over (partition by year order by month asc) as order_count,
from
(SELECT extract(year from order_purchase_timestamp) as year,
order_id,
extract(month from order_purchase_timestamp) as month
 FROM `case-study-1-430318.Target_Dataset.orders`
) X
```

Row	year	month	order_count
1	2016	9	4
2	2016	10	328
3	2016	12	329
4	2017	1	800
5	2017	2	2580
6	2017	3	5262
7	2017	4	7666
8	2017	5	11366
9	2017	6	14611
10	2017	7	18637
11	2017	8	22968
12	2017	9	27253
13	2017	10	31884
14	2017	11	39428
15	2017	12	45101
16	2018	1	7269
17	2018	2	13997

Inference & Insights:

The dataset was sorted by year and month, and the **order count** was calculated for each period.

1. In **2016**, there were **329 orders**, while in **2017**, the count rose sharply to **45,101 orders**.
2. This indicates a **significant growth trajectory**, reflecting either business expansion, increased customer adoption, or improved platform reach.
3. The **gradual month-on-month increase** highlights growing trust and engagement from customers.
4. Such insights can be used to forecast future sales trends, allocate resources efficiently, and design targeted marketing strategies during peak growth phases.

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

Hint: We want you to find out if the no. of orders placed are at peak during certain months.

```
select distinct month_no, month,
count(order_id) over (partition by month_no order by month_no asc) as order_count
from
(SELECT order_id,
extract(year from order_purchase_timestamp) as year,
extract(month from order_purchase_timestamp) as month_no,
FORMAT_DATE("%B", order_purchase_timestamp) AS month,
 FROM `case-study-1-430318.Target_Dataset.orders`) X
order by order_count desc
```

Row	month_no	month	order_count
1	8	August	10843
2	5	May	10573
3	7	July	10318
4	3	March	9893
5	6	June	9412
6	4	April	9343
7	2	February	8508
8	1	January	8069
9	11	November	7544
10	12	December	5674
11	10	October	4959
12	9	September	4305

Another way:

```
select distinct year, month_no, month,
count(order_id) over (partition by year,month_no order by month_no asc) as order_count
from
(SELECT order_id,
extract(year from order_purchase_timestamp) as year,
extract(month from order_purchase_timestamp) as month_no,
FORMAT_DATE("%B", order_purchase_timestamp) AS month,
 FROM `case-study-1-430318.Target_Dataset.orders`) X
```

Row	year	month_no	month	order_count
1	2016	9	September	4
2	2016	10	October	324
3	2016	12	December	1
4	2017	1	January	800
5	2017	2	February	1780
6	2017	3	March	2682
7	2017	4	April	2404
8	2017	5	May	3700
9	2017	6	June	3245
10	2017	7	July	4026
11	2017	8	August	4331
12	2017	9	September	4285
13	2017	10	October	4631
14	2017	11	November	7544
15	2017	12	December	5673
16	2018	1	January	7269
17	2018	2	February	6728

Inference & Insights:

After calculating the number of orders placed each month, it was observed that **August consistently shows the highest order volume**, indicating a clear seasonal peak.

1. The spike in August suggests **increased consumer activity** during this period, which may be linked to seasonal sales, festivals, or marketing campaigns.
2. Businesses can leverage this insight by **strengthening inventory planning, logistics, and promotional strategies** ahead of August to maximize sales.
3. Other months show comparatively lower volumes, highlighting the importance of **targeted campaigns during peak demand periods**.

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

a. **0-6 hrs : Dawn**
b. **7-12 hrs : Mornings**
c. **13-18 hrs : Afternoon**
d. **19-23 hrs : Night**

Hint: We want you to categorize the hours of a day into the given time brackets/ intervals and find out during which intervals the Brazilian customers usually order the most.

```sql
Select time_of_day,
count(distinct order_id) as total_orders
from
(SELECT c.customer_id,o.order_id,
CAST(order_purchase_timestamp AS TIME) AS time,
(case when CAST(order_purchase_timestamp AS TIME) between '00:00:00' and '06:00:00'
then 'Dawn'
when CAST(order_purchase_timestamp AS TIME) between '06:00:00' and '12:00:00'
then 'Mornings'
when CAST(order_purchase_timestamp AS TIME) between '12:00:00' and '18:00:00'
then 'Afternoon'
when CAST(order_purchase_timestamp AS TIME) between '18:00:00' and '23:59:59'
then 'Night'
end ) as time_of_day
FROM `case-study-1-430318.Target_Dataset.customers` c join
`case-study-1-430318.Target_Dataset.orders` o
on c.customer_id = o.customer_id
) X
group by time_of_day
ORDER BY total_orders asc
```

JOB INFORMATION | RESULTS | CHART | JSON

Row	time_of_day	total_orders
1	Dawn	4740
2	Mornings	22240
3	Night	34096
4	Afternoon	38365

Inference & Insights:

By grouping order purchase times into four stages (**Dawn, Morning, Afternoon, and Night**), it was observed that **Brazilian customers most frequently place their orders in the Afternoon**.

1. This reflects **higher digital engagement during mid-day hours**, possibly aligned with lunch breaks or post-work routines.
2. **Morning and Night** show moderate activity, while **Dawn has the lowest order volume**, which aligns with typical customer behavior.
3. Such insights can guide **marketing campaigns and promotions** — e.g., scheduling push notifications, discounts, or ads in the **Afternoon window** to maximize conversions.

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

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

Hint: We want you to get the no. of orders placed in each state, in each month by our customers.

```
Select customer_state, month_num, count(order_id) as order_count
from
(SELECT c.customer_id,c.customer_state,o.order_id,
extract(month from o.order_purchase_timestamp) as month_num,
format_date('%B', o.order_purchase_timestamp) as month
FROM `case-study-1-430318.Target_Dataset.customers` c join
`case-study-1-430318.Target_Dataset.orders` o
on c.customer_id = o.customer_id) X
group by customer_state, month_num
order by customer_state asc, month_num
```

Row	customer_state	month_num	order_count
1	AC	1	8
2	AC	2	6
3	AC	3	4
4	AC	4	9
5	AC	5	10
6	AC	6	7
7	AC	7	9
8	AC	8	7
9	AC	9	5
10	AC	10	6
11	AC	11	5
12	AC	12	5
13	AL	1	39
14	AL	2	39
15	AL	3	40
16	AL	4	51

Inference & Insights:

The data was grouped by **state** and **month**, and the **number of orders** was calculated for each unique combination.

1. This breakdown highlights **state-wise seasonal trends** — for example, larger states like **São Paulo (SP), Rio de Janeiro (RJ), and Minas Gerais (MG)** consistently record the highest monthly order volumes.
2. Smaller states contribute fewer orders, but still show **distinct seasonal peaks**, often aligning with nationwide shopping trends (e.g., year-end sales).
3. This analysis enables **comparison of customer behavior across regions**, helping businesses identify **key markets** and tailor **marketing strategies** by state and season.
4. Month-on-month tracking also provides valuable input for **demand forecasting** and **supply chain optimization** at the state level.

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

Hint: We want you to get the no. of unique customers present in each state.

```
Select customer_state, count(distinct customer_id) as unique_customer_count
from
(SELECT c.customer_id,c.customer_state,o.order_id,
FROM `case-study-1-430318.Target_Dataset.customers` c join
`case-study-1-430318.Target_Dataset.orders` o
on c.customer_id = o.customer_id ) X
group by customer_state
order by customer_state asc
```

Row	customer_state	unique_customer_count
1	AC	81
2	AL	413
3	AM	148
4	AP	68
5	BA	3380
6	CE	1336
7	DF	2140
8	ES	2033
9	GO	2020
10	MA	747
11	MG	11635
12	MS	715
13	MT	907
14	PA	975
15	PB	536
16	PE	1652

Inference & Insights:

The data was grouped by **unique customer states**, and the **distinct customer count** was calculated for each state.

1. It was observed that **Minas Gerais (MG)** has the **highest customer base with 11,635 customers**, making it a key market.
2. In contrast, **Amapá (AP)** has the **lowest customer count at just 68**, reflecting minimal customer penetration.
3. The wide variation in customer distribution across states highlights the importance of **regional strategies** — focusing resources on high-density states like MG, SP, and RJ, while exploring ways to expand customer acquisition in smaller states.
4. This insight helps in **market segmentation, targeted promotions, and demand planning** at the state level.

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).
Hint: You can use the payment_value column in the payments table to get the cost of orders.

```
select year,month_num,
ROUND((( next_month_cost - cost) / cost), 2) * 100 AS percentage
from
(select year, month_num, cost,
lag(cost) over (partition by year order by month_num asc) as next_month_cost
from
(SELECT
extract(year from o.order_purchase_timestamp) as year,
extract(month from o.order_purchase_timestamp) as month_num,
SUM(p.payment_value) AS cost
FROM `case-study-1-430318.Target_Dataset.payments` p join
`case-study-1-430318.Target_Dataset.orders` o
on p.order_id =o.order_id
where (extract(month from o.order_purchase_timestamp) between 0 and 8) and
(extract(year from o.order_purchase_timestamp)) between 2017 and 2018
group by year, month_num
order by year, month_num asc) X) Y
order by year, month_num asc
```

Row	year	month_num	percentage
1	2017	1	null
2	2017	2	-53.0
3	2017	3	-35.0
4	2017	4	8.0
5	2017	5	-30.0
6	2017	6	16.0
7	2017	7	-14.0000000000...
8	2017	8	-12.0
9	2018	1	null
10	2018	2	12.0
11	2018	3	-14.0000000000...
12	2018	4	-0.0
13	2018	5	1.0
14	2018	6	13.0
15	2018	7	-4.0
16	2018	8	4.0

Inference & Insights:

The data was grouped by **year** and **month**, and the **percentage increase in order cost** was calculated using the **LAG function** to compare each month's cost with the previous period.

1. For example, in **June**, there was a **16% increase in order cost from 2017 to 2018**, indicating stronger sales performance.
2. This month-on-month percentage change reveals **growth trends and seasonal fluctuations** in customer spending.
3. Positive percentage increases highlight **expansion periods** where customer demand and order values grew, while negative changes may indicate **seasonal slowdowns or reduced purchasing power**.
4. Such analysis is critical for **forecasting, budgeting, and planning promotional campaigns**, as it helps identify periods of rapid growth or decline.

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

Hint: We want you to fetch the total price and the average price of orders for each state.

```
select customer_state as State,
round(sum(price),2) as Total_price,
round(avg(price),2) as Avg_price
from
(SELECT i.order_id,c.customer_state,i.price
```

```
FROM `case-study-1-430318.Target_Dataset.orders` o join
`case-study-1-430318.Target_Dataset.customers` c
on o.customer_id = c.customer_id
join
`case-study-1-430318.Target_Dataset.order_items` i
on i.order_id = o.order_id
order by c.customer_state asc) X
group by customer_state
order by customer_state asc
```

Row	State	Total_price	Avg_price
1	AC	15982.95	173.73
2	AL	80314.81	180.89
3	AM	22356.84	135.5
4	AP	13474.3	164.32
5	BA	511349.99	134.6
6	CE	227254.71	153.76
7	DF	302603.94	125.77
8	ES	275037.31	121.91
9	GO	294591.95	126.27
10	MA	119648.22	145.2
11	MG	1585308.03	120.75
12	MS	116812.64	142.63
13	MT	156453.53	148.3
14	PA	178947.81	165.69
15	PB	115268.08	191.48
16	PE	262788.03	145.51

Inference & Insights:

The data was grouped by **unique customer states**, and for each state the **total order value** was calculated using the `SUM()` function, while the **average order value** was computed using the `AVG()` function.

1. This analysis shows the **overall revenue contribution** of each state (via total order value), helping identify **high-revenue markets**.
2. The **average order value (AOV)** highlights customer **spending behavior per order** in different states.
3. States with **high totals and high AOV** (e.g., SP, RJ, MG) are both **large and high-value markets**.

4. States with **lower totals but higher AOV** may represent **niche premium customers**, while **low totals and low AOV** suggest underdeveloped markets.
5. These insights can be leveraged for **targeted marketing, localized pricing strategies, and sales prioritization**.

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

Hint: We want you to fetch the total freight value and the average freight value of orders for each state.

```
select customer_state as State,
round(sum(freight_value),2) as Total_freight,
round(avg(freight_value),2) as Avg_freight
from
(SELECT i.order_id,c.customer_state,i.freight_value
FROM `case-study-1-430318.Target_Dataset.orders` o join
`case-study-1-430318.Target_Dataset.customers` c
on o.customer_id = c.customer_id
join
`case-study-1-430318.Target_Dataset.order_items` i
on i.order_id = o.order_id
order by c.customer_state asc) X
group by customer_state
order by customer_state asc
```

Row	State	Total_freight	Avg_freight
1	AC	3686.75	40.07
2	AL	15914.59	35.84
3	AM	5478.89	33.21
4	AP	2788.5	34.01
5	BA	100156.68	26.36
6	CE	48351.59	32.71
7	DF	50625.5	21.04
8	ES	49764.6	22.06
9	GO	53114.98	22.77
10	MA	31523.77	38.26
11	MG	270853.46	20.63
12	MS	19144.03	23.37
13	MT	29715.43	28.17
14	PA	38699.3	35.83

Inference & Insights:

The data was grouped by unique customer states, and for each state the total freight was calculated using the `SUM()` function, while the average freight was computed using the `AVG()` function.

1. This analysis shows the overall freight contribution of each state, helping identify regions with the highest shipping volume.
2. The average freight highlights the cost per order in different states, indicating shipping efficiency or higher logistics costs.
3. States with lower total freight but higher average freight may represent regions with fewer but larger shipments, while low total and low average freight suggest underutilized or smaller markets.
4. These insights can be used to optimize logistics, plan cost-effective shipping strategies, and prioritize high-impact regions.

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.

Hint: 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

```
select *
from
(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 `case-study-1-430318.Target_Dataset.orders` ) X
where time_to_deliver is not NULL and diff_estimated_delivery is not NULL
limit 15
```

Row	order_id	time_to_deliver	diff_estimated_delivery
1	770d331c84e5b214bd9dc70a...	7	45
2	1950d777989f6a877539f5379...	30	-12
3	2c45c33d2f9cb8ff8b1c86cc28...	30	28
4	dabf2b0e35b423f94618bf965f...	7	44
5	8beb59392e21af5eb9547ae1a...	10	41
6	65d1e226dfaeb8cdc42f66542...	35	16
7	c158e9806f85a33877bdfd4f60...	23	9
8	b60b53ad0bb7dacacf2989fe2...	12	-5
9	c830f223aae08493ebecb52f2...	12	12
10	a8aa2cd070eeac7e4368cae3d...	7	1
11	813c55ce9b6baa8f879e064fbf...	12	9
12	44558a1547e448b41c48c4087...	1	5
13	036b791897847cdb8e39df794...	6	0
14	1aba60c04110bdd421b250ea...	21	7
15	0312ecf90786def87f98aa19e0...	7	0

Inference & Insights:

The delivery performance was analyzed by calculating the **time to deliver** for each order based on `order_id`, along with the difference between the estimated and actual delivery dates.

1. For example, for row 1, the actual delivery time was 7 days, but the difference from the estimated delivery date was 45 days, indicating a significant overestimation.
2. This analysis highlights discrepancies between estimated and actual delivery times, helping identify inefficiencies in delivery planning.
3. Insights from this data can be used to improve delivery time estimates, enhance customer satisfaction, and optimize logistics processes.

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

Hint: We want you to find the top 5 & the bottom 5 states arranged in increasing order of the average freight value.

```
(select customer_state as State,
round(avg(freight_value),2) as Avg_price,
'Bottom 5 states' AS states
from
(SELECT i.order_id,c.customer_state,i.freight_value
FROM `case-study-1-430318.Target_Dataset.orders` o join
`case-study-1-430318.Target_Dataset.customers` c
```

```
on o.customer_id = c.customer_id
join
`case-study-1-430318.Target_Dataset.order_items` i
on i.order_id = o.order_id
order by c.customer_state asc) X
group by customer_state
order by Avg_price asc
limit 5)

UNION all

(select customer_state as State,
round(avg(freight_value),2) as Avg_price,
'Top 5 states' AS states
from
(SELECT i.order_id,c.customer_state,i.freight_value
FROM `case-study-1-430318.Target_Dataset.orders` o join
`case-study-1-430318.Target_Dataset.customers` c
on o.customer_id = c.customer_id
join
`case-study-1-430318.Target_Dataset.order_items` i
on i.order_id = o.order_id
order by c.customer_state asc) X
group by customer_state
order by Avg_price desc
limit 5 )
```

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

Inference & Insights:

The analysis was performed by taking a `UNION ALL` of two queries: the first identifies the **Top 5 states** and the second identifies the **Bottom 5 states** based on their **average freight**.

1. This highlights states with the highest and lowest shipping costs, helping to understand regional freight patterns.
2. States with high average freight may indicate higher shipping complexity or longer distances, while low average freight suggests more efficient or shorter deliveries.
3. These insights can guide logistics optimization, cost management, and targeted strategies for regions with high shipping expenses.

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

Hint: We want you to find the top 5 & the bottom 5 states arranged in increasing order of the average delivery time.

```
(select customer_state as State,
round(avg(time_to_deliver),2) as Avg_time_to_deliver,
'Bottom 5 states' AS states
from
(SELECT o.order_id,c.customer_state,
DATE_DIFF(o.order_delivered_customer_date, o.order_purchase_timestamp, DAY) AS
time_to_deliver
FROM `case-study-1-430318.Target_Dataset.orders` o join
`case-study-1-430318.Target_Dataset.customers` c
on o.customer_id = c.customer_id
order by c.customer_state asc) X
group by customer_state
having Avg_time_to_deliver is not NULL
order by Avg_time_to_deliver asc
limit 5)

union all

(select customer_state as State,
round(avg(time_to_deliver),2) as Avg_time_to_deliver,
'Top 5 states' AS states
from
(SELECT o.order_id,c.customer_state,
DATE_DIFF(o.order_delivered_customer_date, o.order_purchase_timestamp, DAY) AS
time_to_deliver
```

```
FROM `case-study-1-430318.Target_Dataset.orders` o join
`case-study-1-430318.Target_Dataset.customers` c
on o.customer_id = c.customer_id
order by c.customer_state asc) X
group by customer_state
having Avg_time_to_deliver is not NULL
order by Avg_time_to_deliver desc
limit 5)
```

Row	State	Avg_time_to_deliver	states
1	RR	28.98	Top 5 states
2	AP	26.73	Top 5 states
3	AM	25.99	Top 5 states
4	AL	24.04	Top 5 states
5	PA	23.32	Top 5 states
6	SP	8.3	Bottom 5 states
7	PR	11.53	Bottom 5 states
8	MG	11.54	Bottom 5 states
9	DF	12.51	Bottom 5 states
10	SC	14.48	Bottom 5 states

Inference & Insights:

The analysis was performed by taking a `UNION ALL` of two queries: the first identifies the **Top 5 states** and the second identifies the **Bottom 5 states** based on their **average delivery time**.

1. This highlights states with the fastest and slowest deliveries, helping to understand regional delivery performance.
2. States with high average delivery times may indicate logistical delays or inefficiencies, while low average delivery times suggest smoother operations.
3. These insights can guide process improvements, targeted operational strategies, and better resource allocation to optimize delivery performance.

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.

Hint: Include only the orders that are already delivered.

```sql
select customer_state as State,
round(avg(fast_delivery)) as difference,
'Top 5 states' AS states
from
(SELECT o.order_id,c.customer_state,
DATE_DIFF(o.order_estimated_delivery_date, o.order_delivered_customer_date, DAY) AS
fast_delivery
FROM `case-study-1-430318.Target_Dataset.orders` o join
`case-study-1-430318.Target_Dataset.customers` c
on o.customer_id = c.customer_id
order by fast_delivery asc) X
group by customer_state
having difference is not NULL
order by difference asc
limit 5
```

Row	State	difference	states
1	AL	8.0	Top 5 states
2	MA	9.0	Top 5 states
3	SE	9.0	Top 5 states
4	SP	10.0	Top 5 states
5	BA	10.0	Top 5 states

Inference & Insights:

The data was grouped by state name, and using the DATE_DIFF function, the difference between the **average actual** and **estimated delivery dates** was calculated to determine delivery speed for each state. The LIMIT function was then used to extract the **Top 5 states** with the fastest deliveries compared to the estimated dates.

1. This analysis identifies states where deliveries consistently outperform estimates, highlighting efficient logistics.
2. States with minimal delivery differences indicate strong operational performance, while larger gaps suggest potential delays.
3. These insights can be leveraged to replicate best practices in slower regions and optimize overall delivery efficiency.

6. Analysis based on the payments:

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

Hint: We want you to count the no. of orders placed using different payment methods in each month over the past years.

```
select payment_type,month_num, count(distinct order_id) as no_of_order
from
(SELECT o.order_id, p.payment_type,
extract(month from o.order_purchase_timestamp) as month_num,
format_date('%B',o.order_purchase_timestamp) as month
FROM `case-study-1-430318.Target_Dataset.payments` p join
`case-study-1-430318.Target_Dataset.orders` o
on p.order_id =o.order_id) X
group by payment_type,month_num
order by payment_type,month_num
```

Row	payment_type	month_num	no_of_order
1	UPI	1	1715
2	UPI	2	1723
3	UPI	3	1942
4	UPI	4	1783
5	UPI	5	2035
6	UPI	6	1807
7	UPI	7	2074
8	UPI	8	2077
9	UPI	9	903
10	UPI	10	1056
11	UPI	11	1509
12	UPI	12	1160
13	credit_card	1	6093
14	credit_card	2	6582
15	credit_card	3	7682
16	credit_card	4	7276
17	credit_card	5	8308

Inference & Insights:

The data was analyzed by counting the number of orders placed using each unique `payment_type` for every month across the past years. For example, the UPI payment type will have order counts for all 12 months, and similarly for other payment types.

1. This analysis highlights the monthly usage trends of different payment methods, helping identify customer preferences.
2. Payment types with consistently high order counts indicate popular and preferred modes among customers.
3. Insights can be used to optimize payment options, offer targeted promotions, and improve the overall customer experience.

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

Hint: We want you to count the no. of orders placed based on the no. of payment installments where at least one installment has been successfully paid.

```
select payment_installments,count(distinct order_id) as no_of_order
from
(SELECT o.order_id, p.payment_installments
FROM `case-study-1-430318.Target_Dataset.payments` p join
`case-study-1-430318.Target_Dataset.orders` o
on p.order_id =o.order_id
where p.payment_installments >= 1)
group by payment_installments
order by payment_installments asc
```

Row	payment_installments	no_of_order
1	1	49060
2	2	12389
3	3	10443
4	4	7088
5	5	5234
6	6	3916
7	7	1623
8	8	4253
9	9	644
10	10	5315
11	11	23
12	12	133
13	13	16
14	14	15
15	15	74
16	16	5
17	17	8

Inference & Insights:
The data was analyzed by counting the **number of orders** based on `payment_installments` where at least one installment has been successfully paid. The data was grouped by the `payment_installment` column, and the total orders were counted for each group.

1. This analysis shows how customers are utilizing installment options and the distribution of orders across different installment plans.
2. Payment plans with higher order counts indicate more popular or preferred installment options.
3. These insights can be used to design flexible payment strategies, promote preferred plans, and increase customer convenience.

Overall Insights:

1. Sales & Revenue Trends:

- The number of orders has grown significantly from 2016 to 2018, reflecting strong business expansion and increasing customer adoption.
- August consistently shows the highest monthly order volume, indicating clear seasonal peaks that may align with festivals, sales, or marketing campaigns.
- States like São Paulo (SP), Rio de Janeiro (RJ), and Minas Gerais (MG) contribute the highest order volumes and revenues, making them key markets.

2. Customer Distribution & Behavior:

- Brazil has 4,119 unique cities across 27 states with varying customer densities. Minas Gerais (MG) has the highest customer base, while Amapá (AP) has minimal penetration.
- Customers most frequently place orders in the **Afternoon**, suggesting high digital engagement during mid-day hours.
- Average order value (AOV) indicates that some states with fewer orders may still represent high-value customers (niche/premium markets).

3. Delivery Performance:

- Delivery times vary significantly across states. Some states consistently deliver faster than estimates, indicating efficient logistics, while others lag behind, highlighting operational inefficiencies.
- Freight analysis shows that high average freight correlates with longer distances or complex deliveries, while low freight indicates efficient shipping or smaller markets.

4. Payment Behavior:

- UPI, credit cards, and other popular payment types dominate, with customers showing clear preferences for certain methods.
- Installment options are widely used, with most customers preferring 1–3 installments, reflecting demand for flexible payment options.

5. Operational Insights:

- States with high total orders and high AOV (SP, RJ, MG) are both large and lucrative, whereas states with low totals and low AOV may require growth strategies.
- Seasonal demand and regional disparities suggest opportunities for better inventory management, marketing, and delivery planning.

Recommendations

1. Market & Customer Strategy:

- Focus marketing campaigns and sales initiatives on high-density and high-value states (SP, RJ, MG) to maximize revenue.
- Develop strategies to increase penetration in underrepresented states like AP and smaller markets through localized promotions, partnerships, or targeted advertising.

2. Delivery & Logistics Optimization:

- Replicate best practices from fast-delivery states to improve slower regions, reducing delivery delays.
- Optimize freight costs by evaluating high-cost regions and exploring alternate shipping routes or carriers.
- Plan inventory and logistics ahead of seasonal peaks (e.g., August) to prevent stockouts and improve customer experience.

3. Payment & Customer Convenience:

- Promote popular payment types and flexible installment options to enhance conversion rates.
- Introduce incentives or discounts for less-used payment methods to balance usage and encourage adoption.

4. Data-Driven Decision Making:

- Use order, freight, and delivery data for predictive analytics to forecast demand, optimize inventory, and allocate resources efficiently.
- Monitor monthly and seasonal trends to plan marketing campaigns, promotional events, and peak-season logistics.

5. Customer Experience Enhancement:

- Ensure accurate delivery estimates to maintain customer trust, especially in states with high delivery discrepancies.
- Gather and analyze reviews to identify product or service issues, enabling improvements in customer satisfaction.