**TARGET SQL CASE STUDY**

**Context:**

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.

The data is available in 8 csv files:

1. customers.csv
2. sellers.csv
3. order\_items.csv
4. geolocation.csv
5. payments.csv
6. reviews.csv
7. orders.csv
8. products.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 **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 **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 **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 **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 **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 **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 |

**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.

**What does 'good' look like?**

1. **Import the dataset and do usual exploratory analysis steps like checking the structure & characteristics of the dataset:**
   1. Data type of all columns in the "customers" table.
   2. Get the time range between which the orders were placed.
   3. Count the Cities & States of customers who ordered during the given period.
2. **In-depth Exploration:**  
   1. Is there a growing trend in the no. of orders placed over the past years?
   2. Can we see some kind of monthly seasonality in terms of the no. of orders being placed?
   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
3. **Evolution of E-commerce orders in the Brazil region:**
   1. Get the month on month no. of orders placed in each state.
   2. How are the customers distributed across all the states?
4. **Impact on Economy: Analyze the money movement by e-commerce by looking at order prices, freight and others.**
   1. Get the % increase in the cost of orders from year 2017 to 2018 (include months between Jan to Aug only).  
      You can use the "payment\_value" column in the payments table to get the cost of orders.
   2. Calculate the Total & Average value of order price for each state.
   3. Calculate the Total & Average value of order freight for each state.
5. **Analysis based on sales, freight and delivery time.**
   1. Find the no. of days taken to deliver each order from the order’s purchase date as delivery time.  
      Also, calculate the difference (in days) between the estimated & actual delivery date of an order.  
      Do this in a single query.  
        
      You can calculate the delivery time and the difference between the estimated & actual delivery date using the given formula:
      * **time\_to\_deliver** = order\_delivered\_customer\_date - order\_purchase\_timestamp
      * **diff\_estimated\_delivery** = order\_delivered\_customer\_date - order\_estimated\_delivery\_date
   2. Find out the top 5 states with the highest & lowest average freight value.
   3. Find out the top 5 states with the highest & lowest average delivery time.
   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.
6. **Analysis based on the payments:**
   1. Find the month on month no. of orders placed using different payment types.
   2. Find the no. of orders placed on the basis of the payment installments that have been paid.

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

Exploratory Data Analysis (EDA):

* There are **96096** unique customers with total **99441** no\_of\_customers

select count(distinct customer\_unique\_id) as Unique\_customers, count(distinct customer\_id) as no\_of\_customers

from `TARGET\_CASE\_STUDY.customers`

* 27 states in Brazil

select distinct geolocation\_state

from `TARGET\_CASE\_STUDY.geolocation`

* According to google, Brazil has **5570 cities** but dataset has duplicate city names with latin/special characters within city column which accompanies total of **8011 cities**
  + Eg, riachao do dantas, riachão do dantas

select distinct geolocation\_city

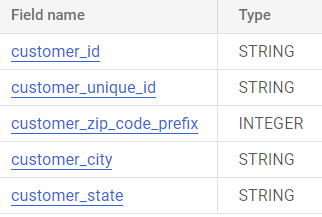
from `TARGET\_CASE\_STUDY.geolocation`

where geolocation\_city like '%riach%'

limit 2

offset 1

1. **Data type of all columns in the "customers" table.**



1. **Get the time range between which the orders were placed.**

select min(order\_purchase\_timestamp) as start\_date, max(order\_purchase\_timestamp) as end\_date

from `TARGET\_CASE\_STUDY.orders`



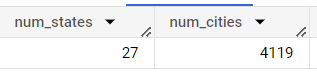
1. **Count the Cities & States of customers who ordered during the given period.**

select count(distinct c.customer\_state) as num\_states, count( distinct c.customer\_city) as num\_cities

from `TARGET\_CASE\_STUDY.customers` c

join `TARGET\_CASE\_STUDY.orders` o on c.customer\_id = o.customer\_id

where o.order\_purchase\_timestamp between '2016-01-01' and '2018-12-31'



**In-depth Exploration:**

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

select T.year, count(T.order\_id) as total\_orders\_per\_year

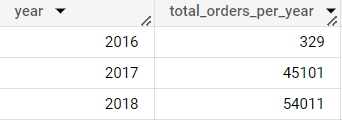
from(

select order\_id, extract(YEAR from order\_purchase\_timestamp) as year

from `TARGET\_CASE\_STUDY.orders`) T

group by 1

order by 1



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

select T.year,T.month,count(T.order\_id) AS orders\_per\_month

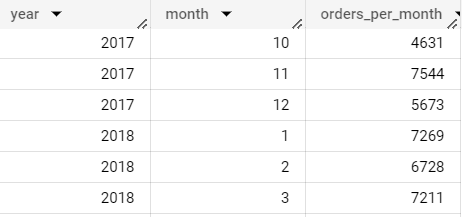
from(

select order\_id, extract(YEAR from order\_purchase\_timestamp) as year,extract(MONTH from order\_purchase\_timestamp) as month

from `TARGET\_CASE\_STUDY.orders`) T

group by 1,2

order by 1,2



* November month has the highest orders in an year

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

select T1.time\_of\_day, count(T1.order\_id) as orders\_at\_particular\_time

from (

select T.\*, case when T.hour between 0 and 6 THEN 'Dawn'

when T.hour between 7 and 12 THEN 'Mornings'

when T.hour between 13 and 18 THEN 'Afternoon'

when T.hour between 19 and 23 THEN 'Night' END as time\_of\_day

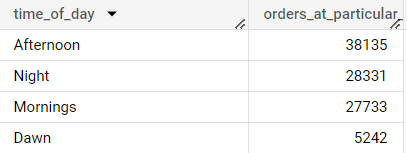
from

( select order\_id, order\_purchase\_timestamp, extract(YEAR from order\_purchase\_timestamp) as year, extract(MONTH from order\_purchase\_timestamp) as month, extract(HOUR from order\_purchase\_timestamp) as hour

from `TARGET\_CASE\_STUDY.orders`) T) T1

group by 1

order by 2 desc



* Most orders are placed during **Afternoon** in an day

**Evolution of E-commerce orders in the Brazil region**

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

select T1.customer\_state as State, concat(T1.YEAR, " - ",T1.month1) as Date, T1.monthly\_orders\_per\_state

from

(select T.customer\_state, T.YEAR,T.month1, count(T.order\_id) as monthly\_orders\_per\_state

From

(select c.customer\_id, o.order\_id, o.order\_purchase\_timestamp, extract(YEAR from order\_purchase\_timestamp) as year, FORMAT\_DATE("%B", o.order\_purchase\_timestamp) as month1, extract(MONTH from order\_purchase\_timestamp) as month, c.customer\_state

from `TARGET\_CASE\_STUDY.customers` c

join `TARGET\_CASE\_STUDY.orders` o on c.customer\_id = o.customer\_id) T

group by 1,2,3,T.month

order by 1,2,T.month

) T1



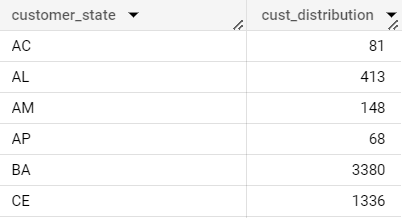
1. How are the customers distributed across all the states?

select customer\_state, count(customer\_id) AS cust\_distribution

from `TARGET\_CASE\_STUDY.customers`

group by 1

order by 1



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

1. Get the % increase in the cost of orders from year 2017 to 2018 (include months between Jan to Aug only).You can use the "payment\_value" column in the payments

select T1.year\_2017, T2.year\_2018, round(((T2.year\_2018 - T1.year\_2017)/ T1.year\_2017 ) \*100,2) as percent\_change

from

(select sum(p.pay\_amount) as year\_2017, row\_number() over() as id --o.order\_id, o.order\_purchase\_timestamp,p.pay\_amount

from `TARGET\_CASE\_STUDY.orders` o

join (select order\_id, sum(payment\_value) as pay\_amount

from `TARGET\_CASE\_STUDY.payments`

group by 1

order by 1) p on o.order\_id = p.order\_id

where o.order\_purchase\_timestamp between '2017-01-01' AND '2017-08-31') T1

join

(select sum(p.pay\_amount) as year\_2018, row\_number() over() as id --o.order\_id, o.order\_purchase\_timestamp,p.pay\_amount

from `TARGET\_CASE\_STUDY.orders` o

join (select order\_id, sum(payment\_value) as pay\_amount

from `TARGET\_CASE\_STUDY.payments`

group by 1

order by 1) p on o.order\_id = p.order\_id

where o.order\_purchase\_timestamp between '2018-01-01' AND '2018-08-31') T2

on T1.id = T2.id



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

select c.customer\_state, round(sum(p.pay\_amount),2) as sum\_of\_price, round(avg(p.pay\_amount),2) as avg\_of\_price --c.customer\_id, c.customer\_state,o.order\_id, o.order\_purchase\_timestamp,p.pay\_amount

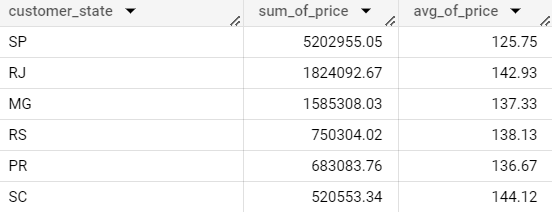
from `TARGET\_CASE\_STUDY.customers` c

join `TARGET\_CASE\_STUDY.orders` o on c.customer\_id = o.customer\_id

join ( select order\_id, sum(price) as pay\_amount from `TARGET\_CASE\_STUDY.order\_items` group by 1) p on o.order\_id = p.order\_id

group by 1

order by 2 desc



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

select c.customer\_state, round(sum(p.pay\_amount),2) as sum\_of\_price, round(avg(p.pay\_amount),2) as avg\_of\_price

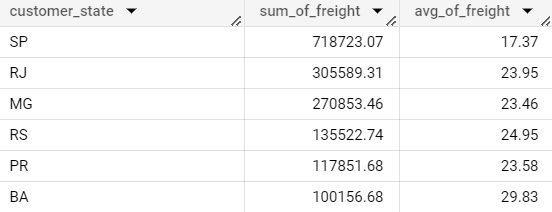
from `TARGET\_CASE\_STUDY.customers` c

join `TARGET\_CASE\_STUDY.orders` o on c.customer\_id = o.customer\_id

join ( select order\_id, sum(freight\_value) as pay\_amount from `TARGET\_CASE\_STUDY.order\_items` group by 1) p on o.order\_id = p.order\_id

group by 1

order by 2 desc



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

1. Find the no. of days taken to deliver each order from the order’s purchase date as delivery time.  
   Also, calculate the difference (in days) between the estimated & actual delivery date of an order.  
   Do this in a single query.  
     
   You can calculate the delivery time and the difference between the estimated & actual delivery date using the given formula:
   1. time\_to\_deliver = order\_delivered\_customer\_date - order\_purchase\_timestamp
   2. diff\_estimated\_delivery = order\_estimated\_delivery\_date - order\_delivered\_customer\_date

select order\_id, customer\_id,order\_status,

FORMAT\_DATE("%d-%m-%Y", order\_purchase\_timestamp) as order\_purchase\_timestamp,

FORMAT\_DATE("%d-%m-%Y",order\_estimated\_delivery\_date) as order\_estimated\_delivery\_date,

FORMAT\_DATE("%d-%m-%Y",order\_delivered\_customer\_date) as order\_delivered\_customer\_date,

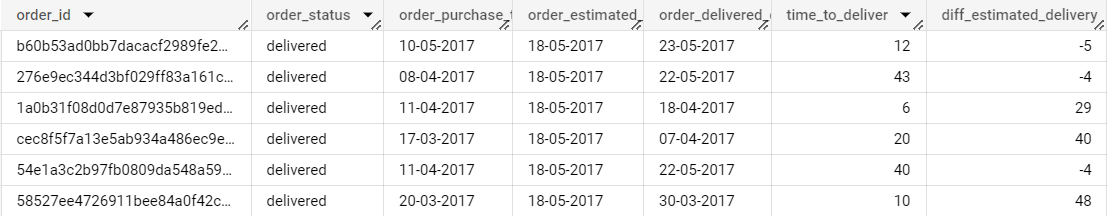
--extract(DAY from order\_purchase\_timestamp) as ordered\_date, extract(DAY from order\_delivered\_customer\_date) as delivered\_date,

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 `TARGET\_CASE\_STUDY.orders`

where order\_status = 'delivered' --or order\_delivered\_customer\_date is null



* Some order\_id has no delivery date, even the order status is ‘delivered’
* Highest no.of days to deliver is 209 days which is 6.8 months

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

select c.customer\_state, round(avg(p.pay\_amount),2) as Highest\_freight\_value

from `TARGET\_CASE\_STUDY.customers` c

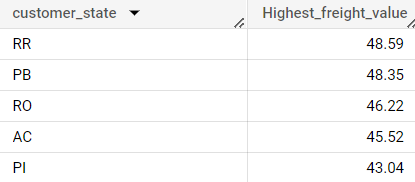
join `TARGET\_CASE\_STUDY.orders` o on c.customer\_id = o.customer\_id

join ( select order\_id, sum(freight\_value) as pay\_amount from `TARGET\_CASE\_STUDY.order\_items` group by 1) p on o.order\_id = p.order\_id

group by 1

order by 2 desc

limit 5



select c.customer\_state, round(avg(p.pay\_amount),2) as lowest\_freight\_value

from `TARGET\_CASE\_STUDY.customers` c

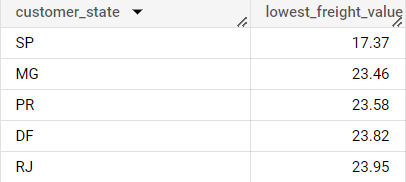
join `TARGET\_CASE\_STUDY.orders` o on c.customer\_id = o.customer\_id

join ( select order\_id, sum(freight\_value) as pay\_amount from `TARGET\_CASE\_STUDY.order\_items` group by 1) p on o.order\_id = p.order\_id

group by 1

order by 2

limit 5



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

select c.customer\_state, round(avg(T.time\_to\_deliver),2) as avg\_deliver\_time

from `TARGET\_CASE\_STUDY.customers` c

join

(select order\_id, customer\_id,order\_status, DATE\_DIFF( order\_delivered\_customer\_date, order\_purchase\_timestamp,DAY) as time\_to\_deliver

from `TARGET\_CASE\_STUDY.orders`

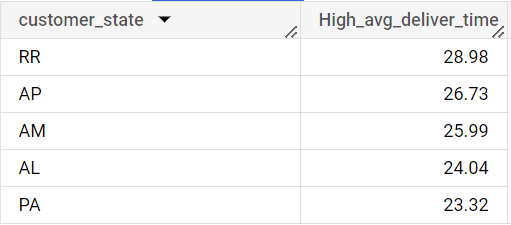
where order\_status = 'delivered') T

on c.customer\_id = T.customer\_id

group by 1

order by 2 desc

limit 5



select c.customer\_state, round(avg(T.time\_to\_deliver),2) as Low\_avg\_deliver\_time

from `TARGET\_CASE\_STUDY.customers` c

Join (select order\_id, customer\_id,order\_status,DATE\_DIFF( order\_delivered\_customer\_date, order\_purchase\_timestamp,DAY) as time\_to\_deliver

from `TARGET\_CASE\_STUDY.orders`

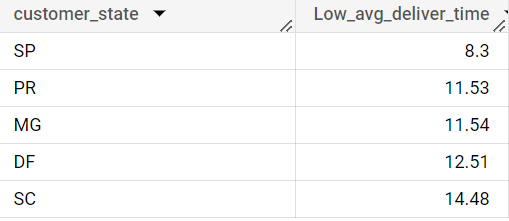
where order\_status = 'delivered') T

on c.customer\_id = T.customer\_id

group by 1

order by 2

limit 5



1. 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.

select c.customer\_state,abs(avg(T.actual\_delivery)- avg(T.estimated\_delivery)) as avg\_delivery\_time

from `TARGET\_CASE\_STUDY.customers` c

join(

select order\_id, customer\_id,order\_status,

DATE\_DIFF( order\_delivered\_customer\_date, order\_purchase\_timestamp,DAY) as actual\_delivery,

DATE\_DIFF(order\_estimated\_delivery\_date, order\_delivered\_customer\_date,DAY) as estimated\_delivery

from `TARGET\_CASE\_STUDY.orders`

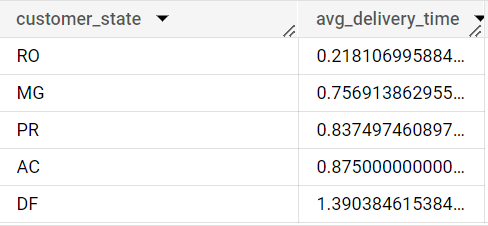
where order\_status = 'delivered') T

on c.customer\_id = T.customer\_id

group by 1

order by 2

limit 5



**Analysis based on the payments**

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

select T.month, T.payment\_type, T.no\_of\_orders

from (

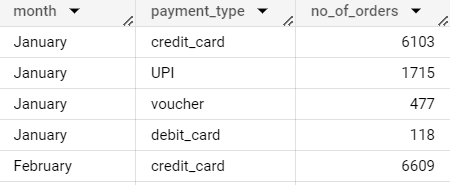
select p.payment\_type, extract(MONTH from o.order\_purchase\_timestamp),FORMAT\_DATE("%B", o.order\_purchase\_timestamp)as month , count(\*) as no\_of\_orders

from `TARGET\_CASE\_STUDY.payments` p

left join `TARGET\_CASE\_STUDY.orders` o on p.order\_id = o.order\_id

group by 1,extract(MONTH from o.order\_purchase\_timestamp),3

order by 2,4 desc) T



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

select p.payment\_installments, count(\*) as orders\_placed

from `TARGET\_CASE\_STUDY.payments` p

left join `TARGET\_CASE\_STUDY.orders` o on p.order\_id = o.order\_id

group by 1

order by 1

