**SQL: “Target” Business Case**

**Table of content**

1. Import the dataset and do usual exploratory analysis steps like checking the structure & characteristics of the dataset
   1. Data type of columns in a table
   2. Time period for which the data is given
   3. Cities and States of customers ordered during the given period
2. In-depth Exploration:
   1. Is there a growing trend on e-commerce in Brazil? How can we describe a complete scenario? Can we see some seasonality with peaks at specific months?
   2. What time do Brazilian customers tend to buy (Dawn, Morning, Afternoon or Night)?
3. Evolution of E-commerce orders in the Brazil region:
   1. Get month on month orders by states
   2. Distribution of customers across the states in Brazil
4. Impact on Economy: Analyze the money movement by e-commerce by looking at order prices, freight and others.
   1. Get % increase in cost of orders from 2017 to 2018 (include months between Jan to Aug only) - You can use “payment\_value” column in payments table
   2. Mean & Sum of price and freight value by customer state
5. Analysis on sales, freight and delivery time
6. Calculate days between purchasing, delivering and estimated delivery
7. Find time\_to\_delivery & diff\_estimated\_delivery. Formula for the same given below:
   * time\_to\_delivery = order\_purchase\_timestamp-order\_delivered\_customer\_date
   * diff\_estimated\_delivery = order\_estimated\_delivery\_date-order\_delivered\_customer\_date
8. Group data by state, take mean of freight\_value, time\_to\_delivery, diff\_estimated\_delivery
9. Sort the data to get the following:
   1. Top 5 states with highest/lowest average freight value - sort in desc/asc limit 5
   2. Top 5 states with highest/lowest average time to delivery
   3. Top 5 states where delivery is really fast/ not so fast compared to estimated date
10. Payment type analysis:
11. Month over Month count of orders for different payment types
12. Count of orders based on the no. of payment instalments

**\*===================\*===================\***

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

* 1. **Data type of columns in a table**
  2. **Time period for which the data is given**
  3. **Cities and States of customers ordered during the given period**

**(1.1) Data type of columns in a table**

**Resources:**

<https://cloud.google.com/bigquery/docs/information-schema-intro> <https://cloud.google.com/bigquery/docs/information-schema-columns>

**Query:**

SELECT

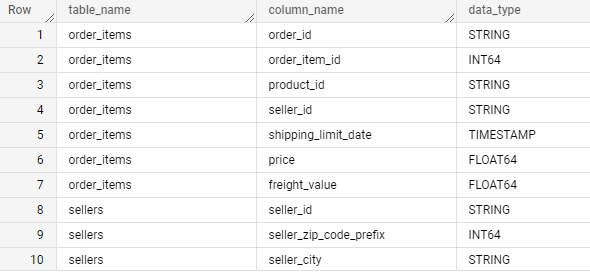
table\_name,

column\_name,

data\_type

FROM Target\_SQL.INFORMATION\_SCHEMA.COLUMNS;

**O/P:**



**Actionable Insights:**

The table provides details about the data types of respective columns belonging to different tables in the dataset.

**(1.2) Time period for which the data is given**

**Query:**

SELECT

  MIN(order\_purchase\_timestamp) AS start\_time\_period\_of\_data,

  MAX(order\_purchase\_timestamp) AS end\_time\_perdiod\_of\_data

FROM `Target\_SQL.orders`;

**O/P:**



**Actionable Insights:**

We see that the “Target” company dataset was provided from the start date of “2016-09-04” until the end data “2018-10-17”

**(1.3) Cities and States of customers ordered during the given period**

**Query:**

SELECT

  DISTINCT

  customer\_state,

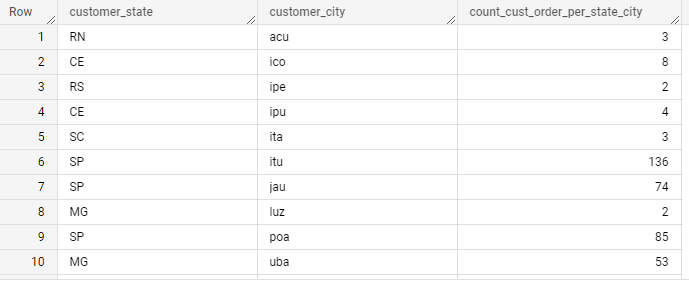
  customer\_city,

  COUNT(customer\_unique\_id) AS count\_cust\_order\_per\_state\_city

FROM `Target\_SQL.customers`

GROUP BY customer\_state, customer\_city;

**O/P:**



**Actionable Insights:**

The table displays the count of orders being placed by the customers in the respective states and cities during the given dataset time period.

**2. In-depth Exploration:**

1. **Is there a growing trend on e\_commerce in Brazil? How can we describe a complete scenario? Can we see some seasonality with peaks at specific months?**
2. **What time do Brazilian customers tend to buy (Dawn, Morning, Afternoon or Night)?**

**(2.1) Is there a growing trend on e\_commerce in Brazil? How can we describe a complete scenario? Can we see some seasonality with peaks at specific months?**

**Query:**

SELECT

  DISTINCT

  COUNT(o.order\_id) AS order\_id,

  EXTRACT(year FROM order\_purchase\_timestamp) AS Year,

  EXTRACT(month FROM order\_purchase\_timestamp) AS Month

FROM `Target\_SQL.orders` as o

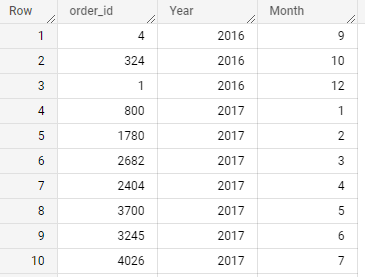
JOIN `Target\_SQL.customers` as c

ON o.customer\_id = c.customer\_id

GROUP BY Year, Month

ORDER BY Year, Month;

**O/P:**



**Actionable Insights:**

As you can see from the table output, there seem to be an increase on the number of orders being placed by the customers as the time passes by starting from the year 2016.

**(2.2) What time do Brazilian customers tend to buy (Dawn, Morning, Afternoon or Night)?**

**Assumptions for this analysis:**

Dawn – 12AM to 7AM

Morning - 8AM to 12PM

Afternoon - 1PM to 6PM

Night - 7PM to 11:59PM

**Query:**

SELECT

  COUNT(order\_id) AS count\_of\_orders,

  CASE

    WHEN EXTRACT(TIME FROM order\_purchase\_timestamp) >= '00:00:01' AND EXTRACT(TIME FROM order\_purchase\_timestamp) < '07:00:00'

    THEN 'Dawn (00.00.01 - 07.00.00)'

    WHEN EXTRACT(TIME FROM order\_purchase\_timestamp) >= '08:00:00' AND EXTRACT(TIME FROM order\_purchase\_timestamp) < '12:00:00'

    THEN 'Morning (08.00.00 - 12.00.00)'

    WHEN EXTRACT(TIME FROM order\_purchase\_timestamp) >= '13:00:00' AND EXTRACT(TIME FROM order\_purchase\_timestamp) < '18:00:00'

    THEN 'Afternoon (13.00.00 - 18.00.00)'

    ELSE 'Night (19.00.00 - 23.00.00)'

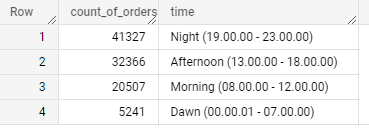
 END AS time

FROM `Target\_SQL.orders`

GROUP BY time

ORDER BY COUNT(order\_id) DESC;

**O/P:**



**Actionable Insights:**

We can see that most of the customers tend to buy/place order during the night time while the least order is placed at the morning time.

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

1. **Get month on month orders by states**
2. **Distribution of customers across the states in Brazil**

**(3.1) Get month on month orders by states**

**Query:**

SELECT

  DISTINCT

  c.customer\_state,

  COUNT(o.order\_id) AS order\_id,

  EXTRACT(year FROM order\_purchase\_timestamp) AS Year,

  EXTRACT(month FROM order\_purchase\_timestamp) AS Month

FROM `Target\_SQL.orders` as o

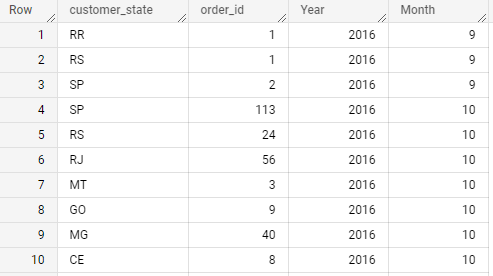
JOIN `Target\_SQL.customers` as c

ON o.customer\_id = c.customer\_id

GROUP BY c.customer\_state, Year, Month

ORDER BY Year, Month;

**O/P:**



**Actionable Insights:**

The table displays the total number of orders placed by customers on monthly bases starting year 2016 until the year 2018 on each state respectively as per the given dataset.

**(3.2) Distribution of customers across the states in Brazil**

**Query:**

SELECT

  DISTINCT

  COUNT(customer\_id) AS customer\_count,

  customer\_state

FROM `Target\_SQL.customers`

GROUP BY customer\_state

ORDER BY customer\_count DESC;

**O/P:**



**Actionable Insights:**

We can see that the highest customers are located on the SP state who places orders in Target marketplace.

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

1. **Get % increase in cost of orders from 2017 to 2018 (include months between Jan to Aug only) - You can use “payment\_value” column in payments table**
2. **Mean & Sum of price and freight value by customer state**

**(4.1)Get % increase in cost of orders from 2017 to 2018 (include months between Jan to Aug only) - You can use “payment\_value” column in payments table**

**Query:**

SELECT

 \*,

 CONCAT(ROUND((x.tot\_pay\_val - (LAG(x.tot\_pay\_val,1) OVER(ORDER BY year, month)))\*100/LAG(x.tot\_pay\_val,1)

    OVER(ORDER BY year,  month),2),"%") AS m\_o\_m\_per100

FROM (SELECT

        EXTRACT(YEAR FROM DATE (order\_purchase\_timestamp)) AS year,

        EXTRACT(MONTH FROM DATE (order\_purchase\_timestamp)) AS month,

        ROUND(SUM(p.payment\_value),2) AS tot\_pay\_val,

        FROM `Target\_SQL.orders` AS o

        JOIN `Target\_SQL.payments` AS p

        ON o.order\_id = p.order\_id

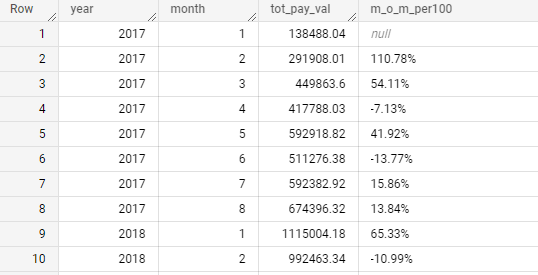
        GROUP BY year,month

        ORDER BY year,month) AS x

WHERE x.year BETWEEN 2017 AND 2018 AND x.month BETWEEN 1 AND 8

ORDER BY year, month;

**O/P:**



**Actionable Insights:**

We can see that the % of cost per orders has been increasing since the beginning of the month January 2017 until the August of 2018 while ignoring the months Sep, Oct, Nov & Dec of 2017.

**(4.2) Mean & Sum of price and freight value by customer state**

**Query:**

SELECT

    c.customer\_state,

    ROUND(AVG(oi.price), 2) AS Avg\_price,

    ROUND(AVG(oi.freight\_value), 2) AS Avg\_freight,

    ROUND(SUM(oi.price), 2) AS Sum\_price,

    ROUND(sum(oi.freight\_value), 2) as Sum\_freight

FROM `Target\_SQL.orders` AS o

JOIN `Target\_SQL.order\_items` AS oi

ON o.order\_id = oi.order\_id

JOIN `Target\_SQL.customers` AS c

ON c.customer\_id = o.customer\_id

GROUP BY c.customer\_state;

**O/P:**



**Actionable Insights:**

As per the table output, we can see that the sum of price and sum of freight is highest in the state SP when compared to other states in Brazil.

1. **Analysis on sales, freight and delivery time**
2. **Calculate days between purchasing, delivering and estimated delivery**
3. **Find time\_to\_delivery & diff\_estimated\_delivery. Formula for the same given below:**
   * **time\_to\_delivery = order\_purchase\_timestamp-order\_delivered\_customer\_date**
   * **diff\_estimated\_delivery = order\_estimated\_delivery\_date-order\_delivered\_customer\_date**
4. **Group data by state, take mean of freight\_value, time\_to\_delivery, diff\_estimated\_delivery**
5. **Sort the data to get the following:**
   1. **Top 5 states with highest/lowest average freight value - sort in desc/asc limit 5**
   2. **Top 5 states with highest/lowest average time to delivery**
   3. **Top 5 states where delivery is really fast/ not so fast compared to estimated date**

**(5.1) Calculate days between purchasing, delivering and estimated delivery**

**Query:**

SELECT

  order\_id,

  order\_purchase\_timestamp,

  order\_delivered\_customer\_date,

  order\_estimated\_delivery\_date,

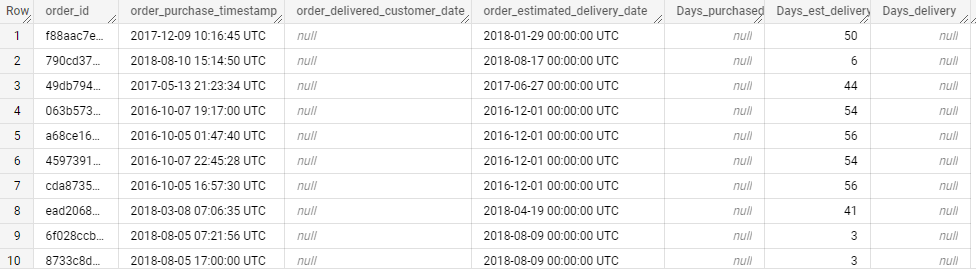
  TIMESTAMP\_DIFF(order\_delivered\_customer\_date, order\_purchase\_timestamp, Day) AS Days\_purchased,

  TIMESTAMP\_DIFF(order\_estimated\_delivery\_date,order\_purchase\_timestamp, Day) AS Days\_est\_delivery,

  TIMESTAMP\_DIFF(order\_estimated\_delivery\_date, order\_delivered\_customer\_date, Day) AS Days\_delivery

FROM `Target\_SQL.orders`;

**O/P:**



**Actionable Insights:**

The table above shows the day’s difference between purchasing, delivering, and estimated delivery days.

**(5.2) Find time\_to\_delivery & diff\_estimated\_delivery. Formula for the same given below:**

* **time\_to\_delivery = order\_purchase\_timestamp-order\_delivered\_customer\_date**
* **diff\_estimated\_delivery = order\_estimated\_delivery\_date-order\_delivered\_customer\_date**

**Query:**

SELECT

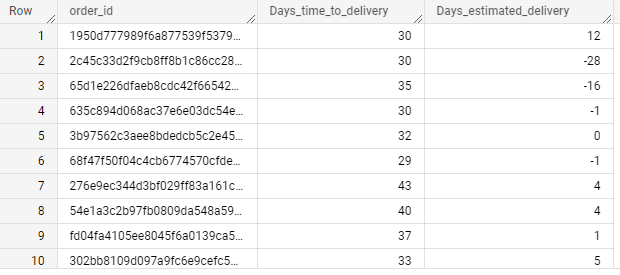
  order\_id,

  TIMESTAMP\_DIFF(order\_delivered\_customer\_date,order\_purchase\_timestamp, Day) AS Days\_time\_to\_delivery,

  TIMESTAMP\_DIFF(order\_delivered\_customer\_date,order\_estimated\_delivery\_date, Day) AS Days\_estimated\_delivery

FROM `Target\_SQL.orders`;

**O/P:**



**Actionable Insights:**

The days’ time to delivery seem to be much longer than the estimated days of delivery in the Brazil country.

**(5.3)Group data by state, take mean of freight\_value, time\_to\_delivery, diff\_estimated\_delivery**

**Query:**

SELECT

  DISTINCT c.customer\_state,

  ROUND(AVG(oi.freight\_value),0) AS Avg\_freight\_value,

  ROUND(AVG(TIMESTAMP\_DIFF(order\_purchase\_timestamp, order\_delivered\_customer\_date, Day)),0) AS Avg\_time\_to\_delivery,

  Round(avg(TIMESTAMP\_DIFF(order\_estimated\_delivery\_date, order\_delivered\_customer\_date, Day)),0) AS Avg\_estimated\_delivery

FROM `Target\_SQL.orders` AS o

JOIN `Target\_SQL.order\_items` AS oi

ON o.order\_id = oi.order\_id

JOIN `Target\_SQL.customers` AS c

ON c.customer\_id = o.customer\_id

GROUP BY c.customer\_state;

**O/P:**



**Actionable Insights:**

The table displays the mean value of freight\_value, time\_to\_delivery and estimated\_delivery on each states respectively.

**(5.4) Sort the data to get the following:**

**(5.4.1) Top 5 states with highest/lowest average freight value - sort in desc/asc limit 5**

**Query:**

SELECT

  c.customer\_state,

  ROUND(AVG(oi.freight\_value),0) AS Highest\_freight\_value

FROM `Target\_SQL.orders` AS o

JOIN `Target\_SQL.order\_items` AS oi

ON o.order\_id = oi.order\_id

JOIN `Target\_SQL.customers` AS c

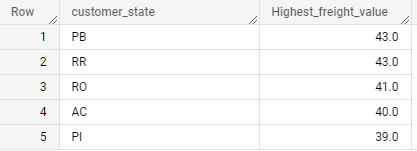
ON c.customer\_id = o.customer\_id

GROUP BY c.customer\_state

ORDER BY Highest\_freight\_value DESC

LIMIT 5;

**O/P:**



**Actionable Insights:**

The top 5 states having the highest average fright value sorted in descending order if reflected on the table output and we see that PB and RR are the two states that’s having the highest freight value of 43.

**(5.4.2) Top 5 states with highest/lowest average time to delivery**

**Query:**

SELECT

  c.customer\_state,

  ROUND(AVG(TIMESTAMP\_DIFF(order\_delivered\_customer\_date,order\_purchase\_timestamp, Day)),0) AS Avg\_time\_to\_delivery,

FROM `Target\_SQL.orders` AS o

JOIN `Target\_SQL.order\_items` AS oi

ON o.order\_id = oi.order\_id

JOIN `Target\_SQL.customers` AS c

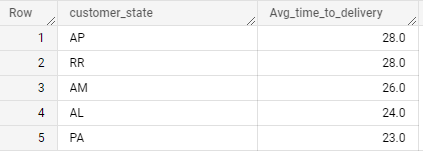
ON c.customer\_id = o.customer\_id

GROUP BY c.customer\_state

ORDER BY Avg\_time\_to\_delivery DESC

LIMIT 5;

**O/P:**



**Actionable Insights:**

As the table output shown above, the states AP and RR are the top 2 states with highest average time to deliver the orders to the customer and then followed by the other 3 states.

**(5.4.3) Top 5 states where delivery is really fast/ not so fast compared to estimated date**

**Query:**

**Fast estimated delivery**

SELECT

  c.customer\_state,

  ROUND(AVG(TIMESTAMP\_DIFF(order\_estimated\_delivery\_date, order\_delivered\_customer\_date, Day)),1) AS fast\_estimated\_delivery

FROM `Target\_SQL.orders` AS o

JOIN `Target\_SQL.order\_items` AS oi

ON o.order\_id = oi.order\_id

JOIN `Target\_SQL.customers` AS c

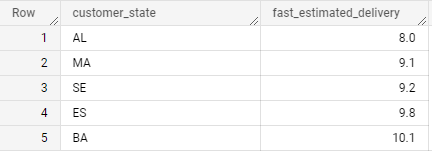
ON c.customer\_id = o.customer\_id

GROUP BY c.customer\_state

ORDER BY fast\_estimated\_delivery ASC

LIMIT 5;

**O/P:**



**Query:**

**Slow estimated delivery**

SELECT

  c.customer\_state,

  ROUND(AVG(TIMESTAMP\_DIFF(order\_estimated\_delivery\_date, order\_delivered\_customer\_date, Day)),1) AS slow\_estimated\_delivery

FROM `Target\_SQL.orders` AS o

JOIN `Target\_SQL.order\_items` AS oi

ON o.order\_id = oi.order\_id

JOIN `Target\_SQL.customers` AS c

ON c.customer\_id = o.customer\_id

GROUP BY c.customer\_state

ORDER BY slow\_estimated\_delivery DESC

LIMIT 5;

**O/P:**



**Actionable Insights:**

The number one state which delivers the orders very fast is AL while on the other side, the number one state which takes the longest time to delivery is AC.

**6. Payment type analysis:**

1. **Month over Month count of orders for different payment types**
2. **Count of orders based on the no. of payment instalments**

**(6.1) Month over Month count of orders for different payment types**

**Query:**

SELECT

  DISTINCT

  EXTRACT(year FROM order\_purchase\_timestamp) AS Year,

  EXTRACT(month FROM order\_purchase\_timestamp) AS Month,

  COUNT(p.order\_id) AS orders, p.payment\_type

FROM `Target\_SQL.payments` AS p

JOIN `Target\_SQL.orders` AS o

ON p.order\_id = o.order\_id

GROUP BY p.payment\_type,Year, Month

ORDER BY Year, Month;

**O/P:**



**Actionable Insights:**

The table displays the count of orders being placed by customers each month based on the different payment type options available. We can also see that most of the customer’s having credit\_card payment type seems to be placing the orders each month when compared to other payment types like voucher or UPI.

**(6.2) Count of orders based on the no. of payment instalments**

**Query:**

SELECT

  COUNT(p.order\_id) AS orders,

  p.payment\_installments

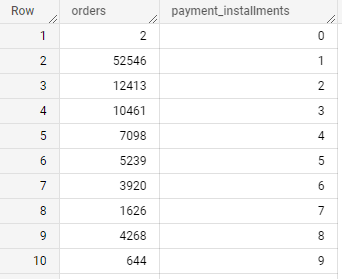
FROM `Target\_SQL.payments` AS p

JOIN `Target\_SQL.orders` AS o

ON p.order\_id = o.order\_id

GROUP BY p.payment\_installments;

**O/P:**



**Actionable Insights:**

As we can see from the table, customers wishing to have payment\_installments of 1 seem to be placed the most number of orders and then followed by 2 and 3 payment instalments options.

**Recommendation:**

"Target" company dataset has been shared from the year 2016 at September 4th until the year 2018 at October 17th along with the various distributions of states and its respective cities.

We can also see that there is a positive growth trend in "Target" since the beginning of the dataset at the year 2016 onwards as majority of the Brazil citizens tend to spend time of placing the orders during the night time between 7PM to 11PM approximately on the E-commerce platform.

To add up on this, since the highest number of customers being placed by the “Target” e-commerce platform is by the SP State, this would again be a very important point that the company can take advantage of that would help to expand the business for obtaining more and more customer’s in that region.

Keeping that aside, I do see that the time gap between the estimated date of arrival and the delivery date seem to be quite high which makes the customer's to wait for a very long time which indeed might end up in having a negative review to the "Target" company from the customer's end.

Since the highest freight\_value is on the PB state and the highest average time to delivery is on the AP state, it would be recommending to double check these sates and see what alternative approaches can be implemented to obtain a faster delivery output to the customer's end.