Data Analysis

1. **Import the dataset and do usual exploratory analysis steps like checking the structure & characteristics of the dataset**
2. Data type of columns in a table

Customers Table Insights: We see customer id is a string or varchar which indicates that it is a combination of letters and numbers. Also, there is another customer unique id which may be generated when the customer made a purchase or it may be their name which is kept private.

Geolocation Table: Geolocation zip code prefix column is common in customers and geolocation which is integer data type. Contains latitude, longitude in float format. City and State in string format.

Order Items: There is order id a string data type which is a unique id for each order placed and order item id an integer data type given to each item within an order. Product id and seller id which are string data type, shipping limit date is timestamp or date time data type and price, fright are float data type.

Order Reviews: There’s a review id a string data type which is unique and generated for each review provided, order id for which order the review is provided, review score an int data type, review creation date a time stamp which seems to be in a wrong format (0001-04-17 00:00:00 UTC), review answer timestamp which is also in the wrong format (0001-04-17 00:00:00 UTC).

Orders : contains order id, customer id, order status in string format. Order purchase, order approved at, delivered to carrier, delivered to customer and estimated delivery timestamp data type columns.

Payments: Order id, payment type are string data type, payment sequential and payment instalments are integers and payment value is a float.

Products: There are few integer data types which corresponds to product description like length, width, height in cm, weight in kg, number of photos of the product in the website and string data types like product id, product category etc

Sellers: There is seller id a string column, seller zip code an integer, seller city and state string data type.

1. Time period for which the data is given

Query:

SELECT extract(date from Min(order\_purchase\_timestamp)) as first\_order\_date,

extract(date from max(order\_purchase\_timestamp)) most\_recent\_order\_date ,

count(order\_id)  total\_number\_orders

FROM `target-dataset123.target\_market.orders`

Result:



Insights: From the above result it’s known that the given data is between 2016 and 2018 with a total close to 100k orders from Brazil during this period.

1. Cities and States of customers ordered during the given period

Query:

select count(distinct customer\_state) total\_number\_of\_states from `target- dataset123.target\_market.customers`

Result:



Insights: Orders came from all the 27 different states in Brazil.

Query:

SELECT distinct customer\_state,count(distinct customer\_city) no\_of\_cities

FROM `target-dataset123.target\_market.customers`

group by customer\_state

order by no\_of\_cities desc

 Result:

Insights: These are the total number of different cities in each state where the customers are from.

1. **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?

Query:

select \*,(number\_of\_orders-lag(number\_of\_orders,1)over(order by Year))/

lag(number\_of\_orders,1)

over(order by Year) groth\_rate

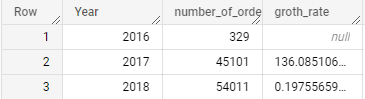
from (SELECT extract(year from order\_purchase\_timestamp) as Year,

count(order\_id) number\_of\_orders

FROM `target-dataset123.target\_market.orders`

group by Year) a

order by Year

 Result:

Insights: As the growth rate is positive, we can say that there is a growing trend on e-commerce in Brazil.

Query:

SELECT extract(year from order\_purchase\_timestamp) as Year,

extract(month from order\_purchase\_timestamp) as Month,

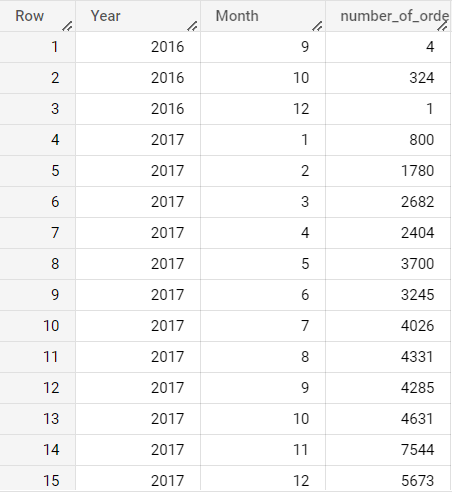
count(order\_id) number\_of\_orders

FROM `target-dataset123.target\_market.orders`

group by Year,Month

order by Year,Month

Result:



Insights: As we have only 3-months of data for the year 2016, only during October the sale was at its peak for that year. In 2017 we see almost a linear growth till the month of October and a sudden exponential raise in the month of November to account for its peak growth for that year. Then during December 2017 there was a decrease in trend but not by much. In the year 2018 the trend stayed almost constant at peak level until the month of August and there was a huge exponential decline in the trend in the month of September and October.

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

Query:

with new\_table as

(select order\_id,Time,case

when Time between ("04:00:00") and ("06:00:00") then "Dawn"

when Time between ("6:00:00" ) and ("12:00:00" )then "Morning"

when Time between ("12:00:00") and ("16:00:00") then "Afternoon"

when Time between ("16:00:00") and ("21:00:00") then "Evening"

else "Night"

end as time\_of\_day

from (select order\_id,extract(time from order\_purchase\_timestamp ) Time

from `target-dataset123.target\_market.orders`) a)

select n1.time\_of\_day,count(n2.time\_of\_day) total\_orders

from new\_table n1

join new\_table n2

on n1.order\_id = n2.order\_id

group by n1.time\_of\_day order by total\_orders desc

Result:



Insights: From the above result we see that 30K (approx. 30%) of the orders were made in the Evening(4 pm UTC to 9 pm UTC), about 25% in the Afternoon(12 pm UTC to 4 pm UTC),22% in the Morning(between 6 am to 12 pm UTC) and approximately 0.4% in the Dawn(9 pm to 4 am UTC).

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

1. Get month on month orders by states

Query:

SELECT c.customer\_state, extract(year from o.order\_purchase\_timestamp) Year,

extract(month from o.order\_purchase\_timestamp) Month,

count(o.order\_id) no\_of\_orders

FROM `target-dataset123.target\_market.orders` o

left join `target-dataset123.target\_market.customers`c

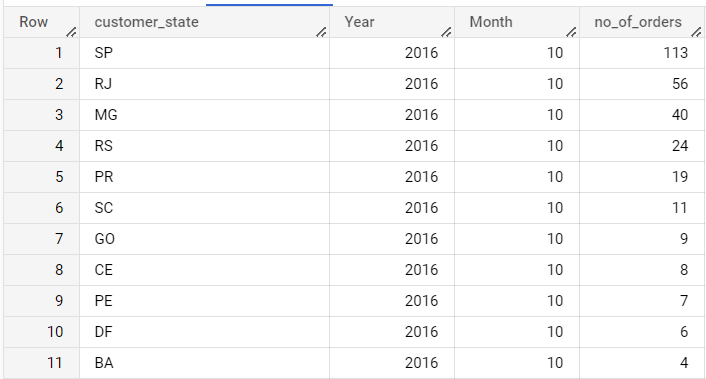
on o.customer\_id = c.customer\_id

group by c.customer\_state,Year,Month

having Year = 2016

order by no\_of\_orders desc

Result:



Insights: São Paulo(SP) had the highest number of orders in October 2016 with 113 orders followed by Rio de Janeiro(RJ) with 56 orders and so on.

Query:

SELECT c.customer\_state, extract(year from o.order\_purchase\_timestamp) Year,

extract(month from o.order\_purchase\_timestamp) Month,

count(o.order\_id) no\_of\_orders

FROM `target-dataset123.target\_market.orders` o

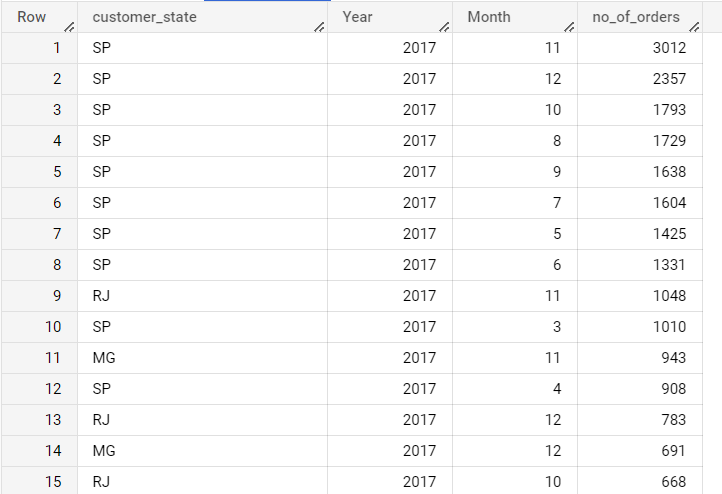
left join `target-dataset123.target\_market.customers`c

on o.customer\_id = c.customer\_id

group by c.customer\_state,Year,Month

having Year = 2017

order by no\_of\_orders desc

Result:

Insights: In 2017 São Paulo(SP) had the highest orders for 8 consecutive months with peak of 3012 orders in November 2017.

Query:

SELECT c.customer\_state, extract(year from o.order\_purchase\_timestamp) Year,

extract(month from o.order\_purchase\_timestamp) Month,

count(o.order\_id) no\_of\_orders

FROM `target-dataset123.target\_market.orders` o

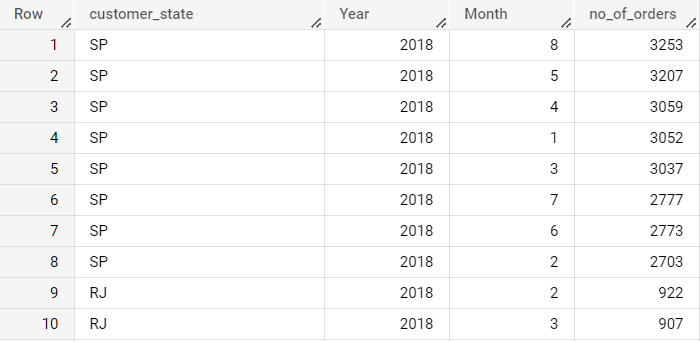
left join `target-dataset123.target\_market.customers`c

on o.customer\_id = c.customer\_id

group by c.customer\_state,Year,Month

having Year = 2018

order by no\_of\_orders desc

Result:

Insights: Again in the year 2018 São Paulo(SP) had the highest orders for 8 consecutive months with peak orders during August 2018 and May 2018 in the second at just 46 orders less than peak

1. Distribution of customers across the states in Brazil

Query:

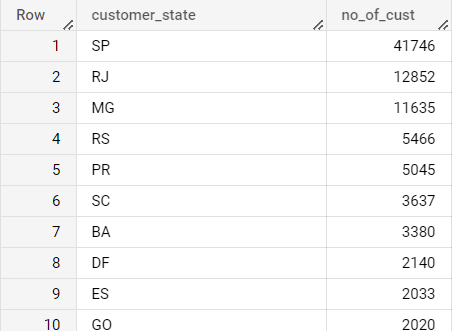
SELECT distinct customer\_state,count(distinct customer\_id) no\_of\_cust

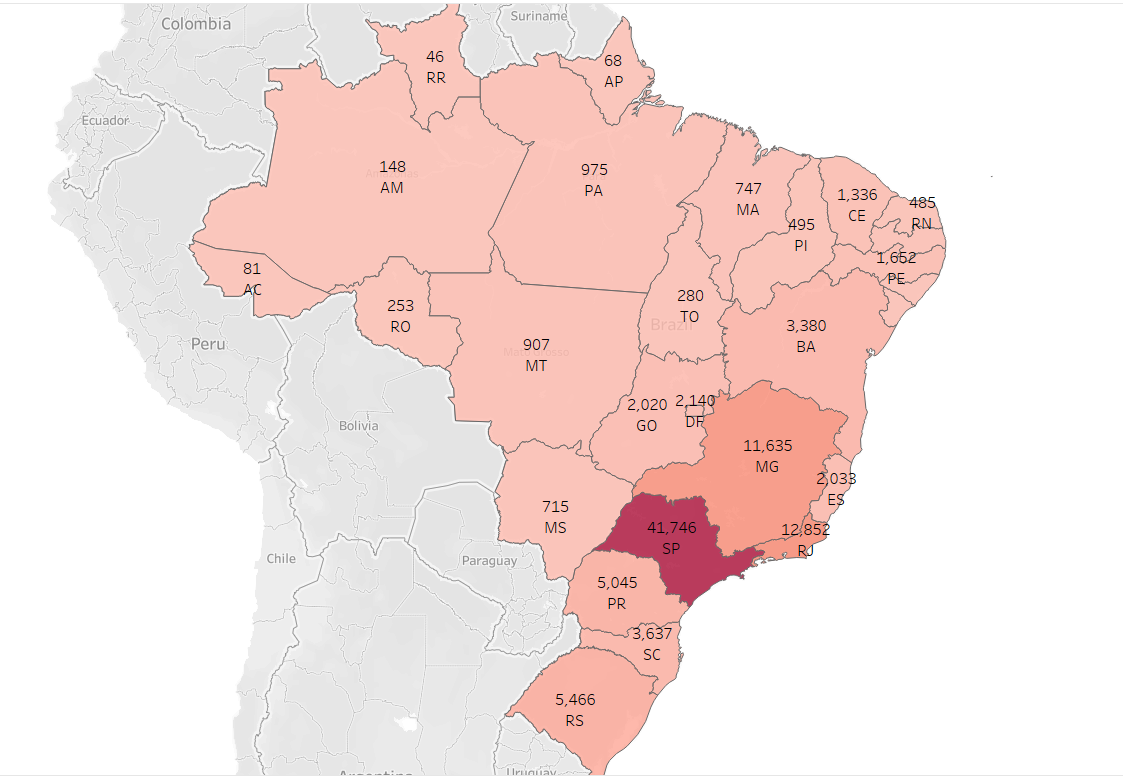
FROM `target-dataset123.target\_market.customers`

group by  customer\_state

order by no\_of\_cust desc

Result:





Insights: SP has most number of customers [41746] that’s the reason why it has more number of orders and RJ with 12,852 customers is in the second spot.

4**. Impact on Economy: Analyse 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

Query:

with table1 as

(SELECT extract(date from o.order\_purchase\_timestamp) Date\_of\_purchase,p.payment\_value FROM `target-dataset123.target\_market.payments` p

left join `target\_market.orders` o on p.order\_id = o.order\_id

where extract(month from o.order\_purchase\_timestamp) between 1 and 8),

table2 as

(select extract(year from Date\_of\_purchase) as Year, sum(payment\_value) as cost\_of\_orders from table1

group by Year

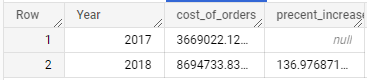
order by Year)

select \*,(cost\_of\_orders -lag(cost\_of\_orders)over(order by Year asc))/

lag(cost\_of\_orders)over(order by Year asc )\*100 as precent\_increase\_cost

from table2

Result:



Insights: There is a significant increase of 137 percent in cost of orders from 2017 to 2018 considering data for the month between Jan to Aug only.

1. Mean & Sum of price and freight value by customer state

Query:

with table1 as

(SELECT oi.order\_id,o.customer\_id,price,freight\_value FROM `target- dataset123.target\_market.order\_items` oi

left join `target\_market.orders` o on o.order\_id = oi.order\_id )

select distinct c.customer\_state,avg(t1.freight\_value)as mean\_freight,

sum(t1.freight\_value) as sum\_freight ,

avg(t1.price) as mean\_price ,

sum(t1.price) as sum\_price

from table1 t1

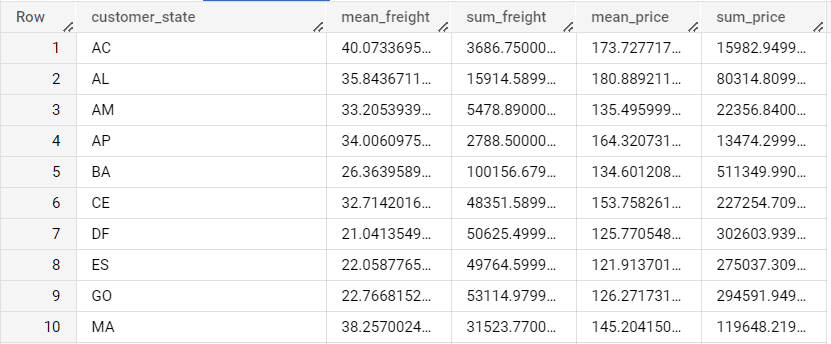
join target\_market.customers c on

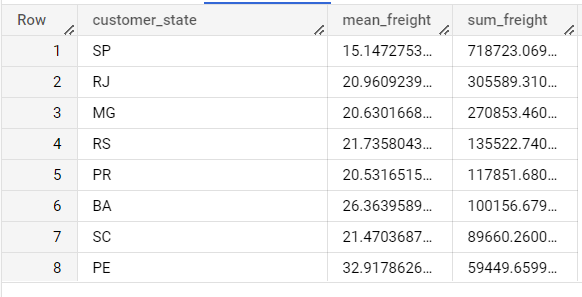
t1.customer\_id = c.customer\_id

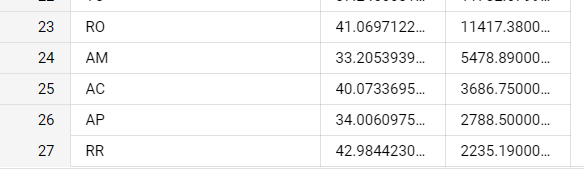
group by c.customer\_state

order by c.customer\_state

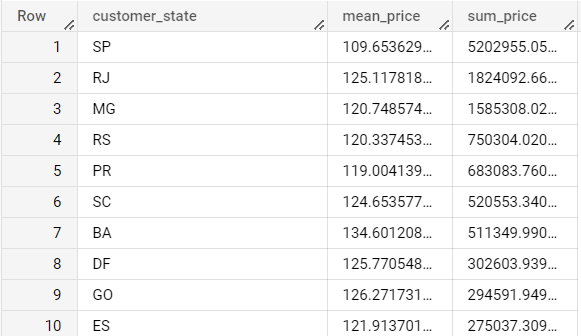
Result:

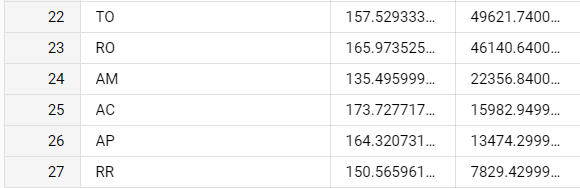






Insights: SP might have highest total freight value but it has the lowest average fright value. This may be due to the highest number of orders from that state. RR(Roraima) has the lowest sum of freight with highest average due to low number of orders.





Insights: Highest sum of prices was seen in SP as it has most number of orders and customers but highest mean sum price was from PB(Paraíba).

5. Analysis on sales, freight and delivery time

1. Calculate days between purchasing, delivering and estimated delivery
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,order\_purchase\_timestamp,order\_delivered\_customer\_date,

date\_diff(order\_delivered\_customer\_date, order\_purchase\_timestamp, day) time\_to\_delivery ,

date\_diff(order\_estimated\_delivery\_date, order\_delivered\_customer\_date, day) diff\_estimate d\_delivery

FROM `target-dataset123.target\_market.orders`

where order\_delivered\_customer\_date is not null and order\_status = "delivered"

order by time\_to\_delivery desc

1. Group data by state, take mean of freight\_value, time\_to\_delivery, diff\_estimated\_delivery

with table1 as

(SELECT o.order\_id,o.order\_purchase\_timestamp,order\_delivered\_customer\_date,

date\_diff(order\_delivered\_customer\_date, order\_purchase\_timestamp, day) time\_to\_delivery ,date\_diff(order\_estimated\_delivery\_date, order\_delivered\_customer\_date, day) diff\_estimated\_delivery ,customer\_state,freight\_value

FROM `target-dataset123.target\_market.orders` o

left join `target\_market.order\_items` oi

on o.order\_id = oi.order\_id

left join `target\_market.customers`c on

o.customer\_id = c.customer\_id

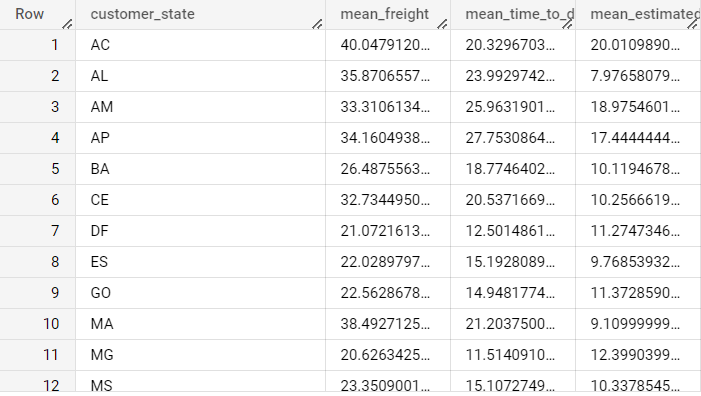
where order\_delivered\_customer\_date is not null and order\_status = "delivered"

order by time\_to\_delivery desc)

select customer\_state,avg(freight\_value) mean\_freight,avg(table1.time\_to\_delivery)mean\_time\_to\_delivery,avg(table1.diff\_estimated\_delivery)mean\_estimated\_delivery from table1

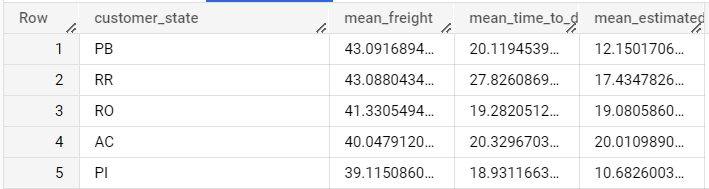
group by customer\_state

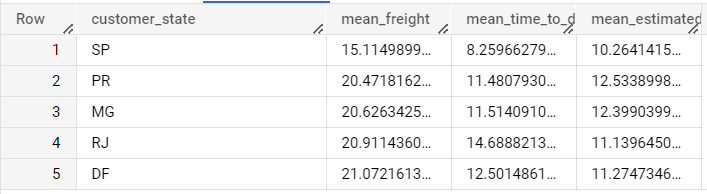
order by customer\_state



1. Sort the data to get the following:
2. Top 5 states with highest/lowest average freight value - sort in desc/asc limit 5

Top 5 highest freight value in descending order



Top 5 lowest freight value in ascending order

Insight: PB and RR has the highest freight value. Some states like PB, RR, RO, AC and PI etc. need to evaluate their method of transportation. They could find a cheaper way of transport. Ship more products at a time, less often. SP has the lowest mean freight charges which is comparatively good which may be due to high number of orders.

1. Top 5 states with highest/lowest average time to delivery

Query: Highest average time.

with table1 as

(SELECT o.order\_id,o.order\_purchase\_timestamp,order\_delivered\_customer\_date,

date\_diff(order\_delivered\_customer\_date,order\_purchase\_timestamp,day)time\_to\_delivery,

date\_diff(order\_estimated\_delivery\_date,order\_delivered\_customer\_date,day)

diff\_estimated\_delivery,customer\_state,freight\_value

FROM `target-dataset123.target\_market.orders` o

left join `target\_market.order\_items` oi

on o.order\_id = oi.order\_id

left join `target\_market.customers`c on

o.customer\_id = c.customer\_id

where order\_delivered\_customer\_date is not null and order\_status = "delivered"

order by time\_to\_delivery desc)

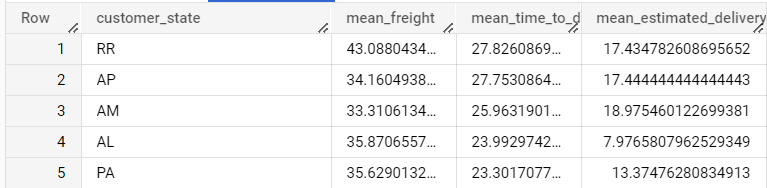
select customer\_state,avg(freight\_value) mean\_freight,avg(table1.time\_to\_delivery)mean\_time\_to\_delivery,avg(table1.diff\_estimated\_delivery)mean\_estimated\_delivery from table1

group by customer\_state

order by mean\_time\_to\_delivery desc

limit 5

Result:

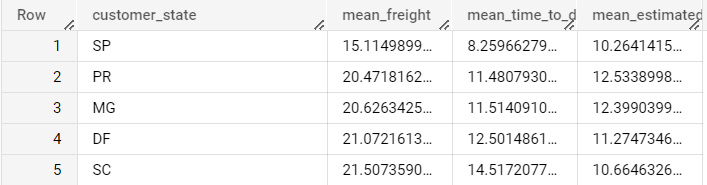


Insights: RR has the highest mean time to delivery is the highest along with AP. This may be because these places are located very far from the product shipping station or it might be a remote place as most parts of RR and AP are covered by Amazon Rain Forest. This can be overcome by setting up small intermediate shipping facilities in these states.

Query: lowest mean time to delivery

order by mean\_time\_to\_delivery asc

limit 5



Insights: SP is the state where mean delivery time is faster than any other state. This is the state which has most number of orders, customers and contributes to most part of the company’s revenue.

1. Top 5 states where delivery is really fast/ not so fast compared to estimated date

Query: Delivery really fast

with table1 as

(SELECT o.order\_id,o.order\_purchase\_timestamp,order\_delivered\_customer\_date,

date\_diff(order\_delivered\_customer\_date,order\_purchase\_timestamp,day)time\_to\_delivery,

date\_diff(order\_estimated\_delivery\_date,order\_delivered\_customer\_date,day)

diff\_estimated\_delivery,customer\_state,freight\_value

FROM `target-dataset123.target\_market.orders` o

left join `target\_market.order\_items` oi

on o.order\_id = oi.order\_id

left join `target\_market.customers`c on

o.customer\_id = c.customer\_id

where order\_delivered\_customer\_date is not null and order\_status = "delivered"

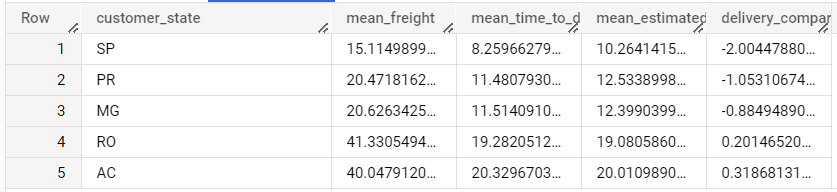
order by time\_to\_delivery desc)

select customer\_state,avg(freight\_value) mean\_freight,avg(table1.time\_to\_delivery)mean\_tim e\_to\_delivery,avg(table1.diff\_estimated\_delivery)mean\_estimated\_delivery,avg(table1.time\_t o\_delivery)-avg(table1.diff\_estimated\_delivery) delivery\_compared\_to\_estimate from table1

group by customer\_state

order by delivery\_compared\_to\_estimate asc

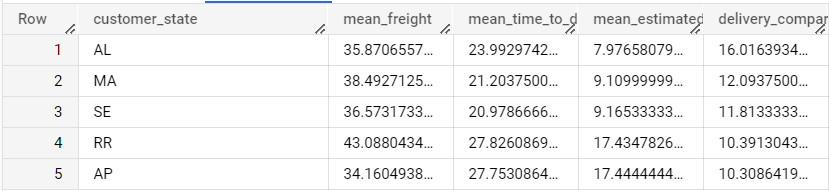
limit 5



Insights: In SP the products were delivered 2 days before the estimated time. Which is good considering the sales in that state. Followed by PR with one day before the estimate, MG, RO and AC with no delay from the estimated date.

Query: Delivery not so fast compared to estimate

order by delivery\_compared\_to\_estimate desc

limit 5

Insights: AL delays there delivery by an average of 16 days.

6. Payment type analysis:

1. Month over Month count of orders for different payment types

Query:

with table1 as

(SELECT p.\*,extract(date from order\_purchase\_timestamp) order\_date FROM `target- dataset123.target\_market.payments` p

left join `target\_market.orders`o

on p.order\_id=o.order\_id)

select distinct payment\_type,extract(month from order\_date)Month,

count(order\_id) over(partition by payment\_type

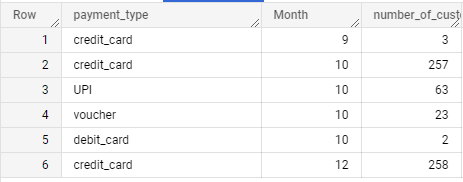
order by extract(month from order\_date)) number\_of\_customers

from table1

where extract(year from order\_date)=2016

order by month asc, number\_of\_customers desc

Result:

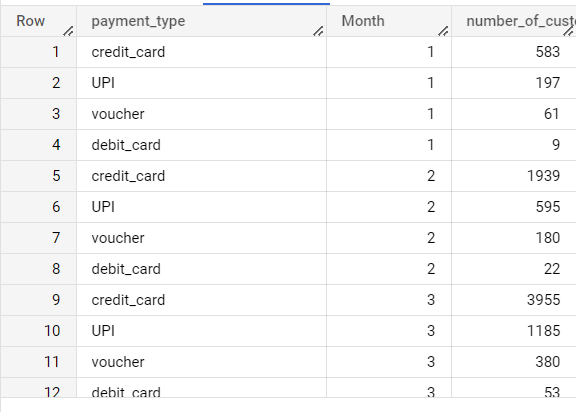


Insights: For the year 2016 all 3 customers used credit card as their mode of payment in September. In October 257 used credit card and 63 used UPI (or it might be some kind of online payment, as UPI is payment mode of India) and few used vouchers. In December all transaction were done by using credit card.

Query:

where extract(year from order\_date)=2017

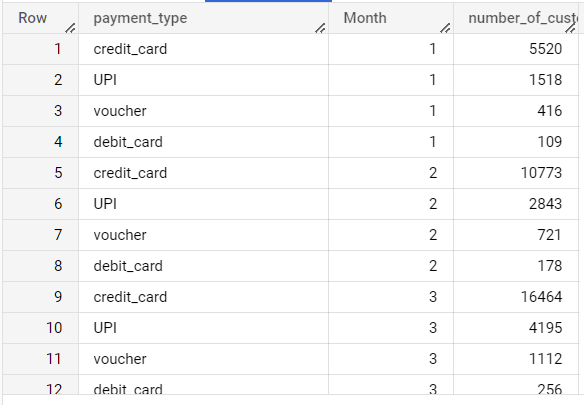
order by month asc, number\_of\_customers desc

Result:

Query:

where extract(year from order\_date)=2018

order by month asc, number\_of\_customers desc

Result:

Insights: In each year and in almost every month the most used payment type is credit card as it is the most convenient way of payment as it is easy and secure. Second most preferred payment mode was UPI? Or online payment. Least used payment mode turns out to be debit card.

2.Count of orders based on the no. of payment instalments

Query:

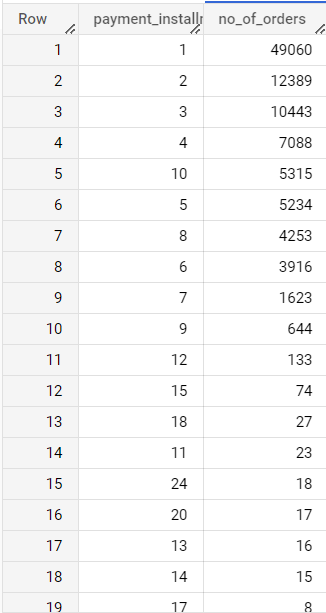
SELECT distinct payment\_installments,count(distinct order\_id)

over(partition by payment\_installments) no\_of\_orders

FROM `target- dataset123.target\_market.payments`

order by no\_of\_orders desc

Result:



Insights: Most orders were paid in a single instalment. Highest number of instalments taken was 24. By this we can assume that most of the customers do not prefer to take loan or they make only affordable purchases.

7.Product category analysis:

1. Total number of different product category and number of orders in each category.

Query:

select count(distinct product\_category) no\_of\_categories

from `target\_market.products`

 Result:

Insights: There are 73 different product categories sold on the website.

Query:

SELECT product\_category,

count(distinct order\_id) no\_of\_orders

FROM `target-dataset123.target\_market.order\_items` o

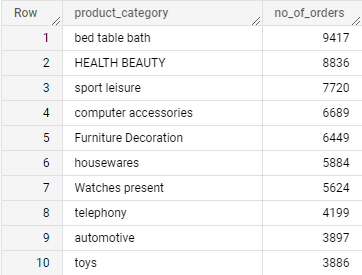
left join `target\_market.products`p

on o.product\_id = p.product\_id

group by product\_category

order by no\_of\_orders desc

Result: Top 10 Bottom 10



Insights: Bed, table and bath category has the highest order count followed by beauty, sport leisure and so on. Insurance and service , PC games and Children’s clothing are the least placed orders.

2. Top 5 and bottom 5 rated products.

Query: Top 5

SELECT product\_category,

round(avg( orev.review\_score),2) mean\_review\_score

FROM `target-dataset123.target\_market.order\_items` oi

join `target\_market.products`p

on oi.product\_id = p.product\_id

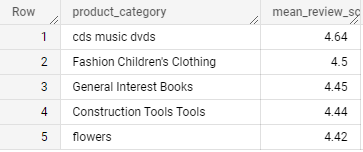
join `target\_market.order\_reviews`orev

on oi.order\_id = orev.order\_id

group by product\_category

order by mean\_review\_score desc

limit 5

 Result

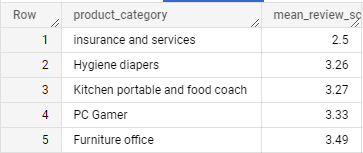
Insights: We can see cds music dvds have been rated and children’s clothing are among the top 5 highest rated products because very few people who purchased those products and most of them have been satisfied that’s the reason they have good rating.

Query:

order by mean\_review\_score asc

limit 5

Result:



Insights: Insurance and services is the lowest rated even though only 2 orders were placed in this category.

This might be one of the reasons for their poor sales.

1. Seller Analysis:
2. Total number of sellers and the Seller who have had highest sales.

Query:

select count(distinct seller\_id) no\_of\_sellers

from `target\_market.sellers`

Result:



Query:

SELECT oi.seller\_id,count(distinct order\_id) no\_of\_orders

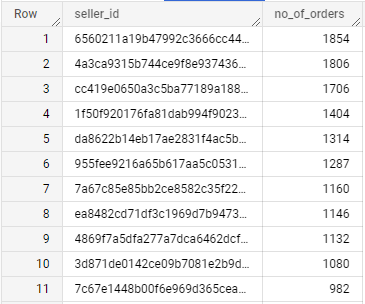
FROM `target-dataset123.target\_market.order\_items` oi

join `target\_market.sellers` s

on oi.seller\_id = s.seller\_id

group by oi.seller\_id

order by no\_of\_orders desc

 Result:

Insights: As there’s no name of the seller these are the sellers who have highest number of orders.

Final Insights:

As this is a vast data set we can do numerous Exploratory data analysis. By doing the necessary analysis we can conclude that state SP is the best performing state in terms of sales and acquiring/ maintaining customer base with minimal freight cost and lowest time to delivery. Whereas State like RR and few other states have least contribution towards sale and have highest freight cost and poor judgement of estimated delivery time.

The poor performing states like RR, AP, AL must consider revising their estimated delivery date and minimize freight charges by shipping more products, less frequently. More freight charges does not mean they are not doing a great job, it may be that these places are too far from where the products are shipped, So, establishing intermediate shipping facilities closer to these far away places might reduce the time to delivery and also the freight cost. This may also result in more order and customers in the future.