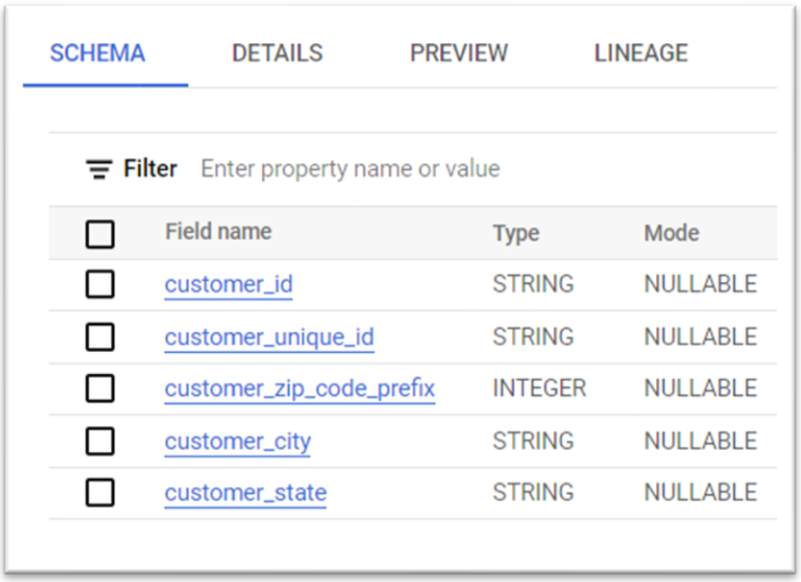
Business Case: Target SQL – SCALER PROJECT

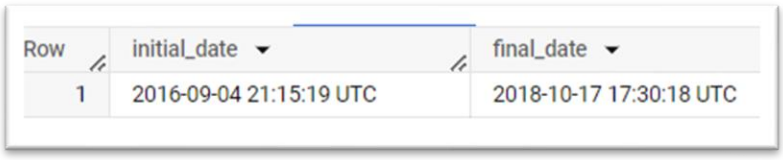
1. Import the dataset and do usual exploratory analysis steps like checking the structure & characteristics of the dataset

1. Data type of columns in a table:

To get the data types of columns of a particular table, click on the table and select schema where we have column name and their respective data-types.



2. Time period for which the data is given



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

```
select extract(year from order_purchase_timestamp) as year,
extract(month from order_purchase_timestamp) as month
from `america_retailer_business.orders`
order by 1,2
```

Insights :
Time period of well-established Brazilian e-commerce company is shown in the above table in the form of timestamp.

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

Row	customer_state	customer_city
1	AC	brasileia
2	AC	cruzeiro do sul
3	AC	epitaciolandia
4	AC	manoel urbano
5	AC	porto acre
6	AC	rio branco
7	AC	senador guiomard
8	AC	xapuri
9	AL	agua branca
10	AL	anadia

Insights:

- Above table shows the country and their respective cities which are ordered in the

lexographical order.

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?

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

```
select extract(year from order_purchase_timestamp) as year,extract(month from order_purchase_timestamp) as month,count(*) as no_of_orders_per_month

from `america_retailer_business.orders`
group by  1,2
order by 1,2;
```

Orders count based on monthly basis irrespective of year

Row	month	no_of_orders_per_m
1	8	10843
2	5	10573
3	7	10318
4	3	9893
5	6	9412
6	4	9343
7	2	8508
8	1	8069
9	11	7544
10	12	5674

```
select * from
(select extract(month from order_purchase_timestamp) as month,count(*) as no_of_orders_per_month
from `america_retailer_business.orders`
group by  extract(month from order_purchase_timestamp))
order by no_of_orders_per_month desc
```

INSIGHTS:

- We can clearly observe that months of May, June, July, August has the highest number of orders.

- Orders in July, August are high because of winter where customers are so comfortable in home.
- Sales in May, June are very high because Brazilian people because of holidays.
- Number of orders in January are also good that's because of New-year celebrations where people exchange their gifts
- In February, March we can observe a spike that's because of Valentin's day and starting of summer season.
- If we observe the table with count of orders based on month-wise for a year, we can clearly observe that there is tremendous growth in the count of orders, which we can infer that in Brazil future of E-Commerce is very high in demand which can become a good business to explore in it.

Recommendations:

- During the months where count of orders were less, we can try to increase company's reputation by adopting marketing strategies and creating campaigns in malls and public places.
- We can plan like summer, winter sales during the period of winter, where customers are very much interested in shopping and stuff.

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

Row	part_of_day	no_of_orders_per_pa
1	Afternoon	38361
2	Night	34100
3	Morning	22240
4	Dawn	4740

```
SELECT
  CASE
    WHEN EXTRACT(hour FROM order_purchase_timestamp) BETWEEN 0 AND 5
THEN "Dawn"
    WHEN EXTRACT(hour FROM order_purchase_timestamp) BETWEEN 6 AND
11 THEN "Morning"
    WHEN EXTRACT(hour FROM order_purchase_timestamp) BETWEEN 12 AND
17 THEN "Afternoon"
    ELSE "Night"
  END AS part_of_day,
  COUNT(*) as no_of_orders_per_part_of_day
```

```
FROM
  `america_retailer_business.orders`
GROUP BY
  1
  order by 2 desc;
```

INSIGHTS:

- It is clearly observed that during afternoon and night we can see a lot number of orders were placed during a part of day.
- Obviously, orders at night are high because people come from work and tend to buy something from sitting in a peaceful position.
- Number of orders in afternoon are very high because people might be having a lunch, nap and leisure period.

Recommendations:

- We can increase customer-care facility during afternoon and night periods.
- We need to maintain a proper inventory so that customers can order their product of choice.

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

1. Get month on month orders by states

Row	customer_state	order_year	order_month	no_of_orders_per_m	month_on_month_sa
1	AC	2017	1	2	2
2	AC	2017	2	3	1
3	AC	2017	3	2	-1
4	AC	2017	4	5	3
5	AC	2017	5	8	3
6	AC	2017	6	4	-4
7	AC	2017	7	5	1
8	AC	2017	8	4	-1
9	AC	2017	9	5	1
10	AC	2017	10	6	1

```
WITH monthly_sales AS (
SELECT
  customer_state,
  EXTRACT(YEAR FROM order_purchase_timestamp) AS order_year,
  EXTRACT(MONTH FROM order_purchase_timestamp) AS order_month,
  COUNT(*) AS no_of_orders_per_month_per_year_state_wise
```

```
  FROM
    `america_retailer_business.orders` o
  JOIN
    `america_retailer_business.customers` c
  ON
    o.customer_id = c.customer_id
  GROUP BY
    1, 2, 3
)
SELECT
  customer_state,
  order_year,
  order_month,
  no_of_orders_per_month_per_year_state_wise,
  IFNULL(no_of_orders_per_month_per_year_state_wise -
LAG(no_of_orders_per_month_per_year_state_wise) OVER (PARTITION BY
customer_state, order_year ORDER BY order_year, order_month),
no_of_orders_per_month_per_year_state_wise) AS month_on_month_sales
FROM
  monthly_sales
ORDER BY
  customer_state, order_year, order_month;
```

Insights:

States like SP, RJ, MG are high in count of orders compare to other countries.

Recommendations:

As we can clearly observe what are states who are interested in e-commerce platforms, we can target them by maintaining inventory and following marketing strategies.

2.Distribution of customers across the states in Brazil

Row	customer_state	count_of_customers
1	SP	41746
2	RJ	12852
3	MG	11635
4	RS	5466
5	PR	5045
6	SC	3637
7	BA	3380
8	DF	2140
9	ES	2033
10	GO	2020

```
select customer_state, count(*) as no_of_customers_per_state

from `america_retailer_business.customers`
group by 1
order by 2 desc
```

Insights:

- Compare to other states, states like SP, RJ and MG have the highest number of customers
- States like AC, AP, RR, RO a has a smaller number of customers compare to other states

Recommendations:

- We can perform market research on states like AC, AP, RR, RO where count of customers is less in number and proceed according to them to target that state customers also.
- Distribution centres, inventory in godowns can be increased as much as required in states like SP, RJ and MG

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

Row	payment_values_on	payment_values_on	percentage_increase
1	3669022.12	8694733.84	136.98

```
with payment_valuess_on_2017 as
(select round(sum(payment_value),2) as payment_values_on_2017
from `america_retailer_business.payments` p
join `america_retailer_business.orders` o
on p.order_id = o.order_id
where extract(year from order_purchase_timestamp) = 2017
and extract(month from order_purchase_timestamp)
between 1 and 8)
,
payment_valuess_on_2018 as
(select round(sum(payment_value),2) as payment_values_on_2018
from `america_retailer_business.payments` p
```

```
join `america_retailer_business.orders` o
on p.order_id = o.order_id
where extract(year from order_purchase_timestamp) = 2018
and extract(month from order_purchase_timestamp)
between 1 and 8)

SELECT
  payment_values_on_2017,
  payment_values_on_2018,
  round((((payment_values_on_2018 - payment_values_on_2017) /
payment_values_on_2017) * 100),2) AS percentage_increase
FROM payment_valuess_on_2017, payment_valuess_on_2018;
```

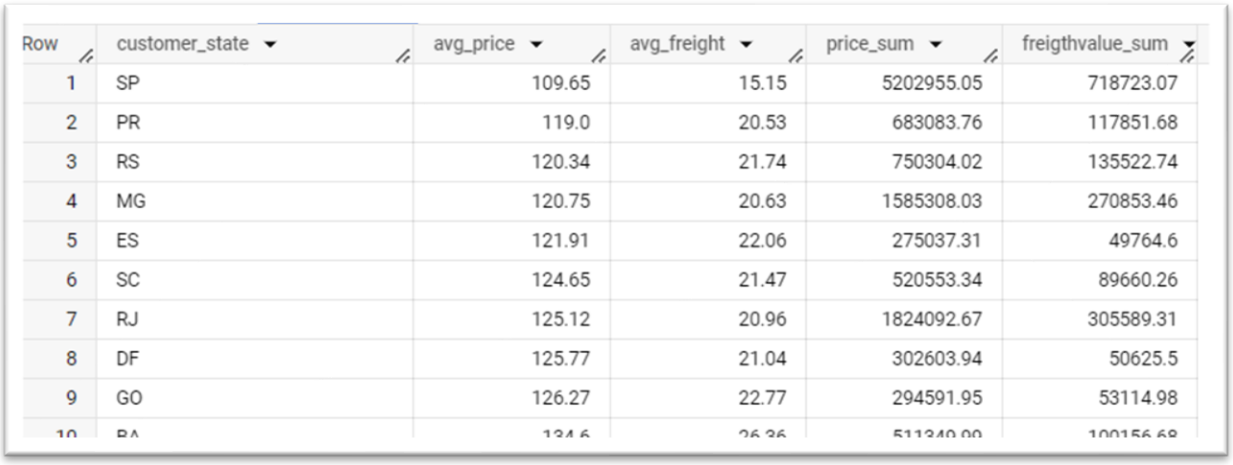
Insights:

- It is clearly observed that, there is a hike of approximately 137% in the cost of orders from the year 2017 to 2018.
- Observation we can infer that future of E-Commerce in Brazil is high in demand and customers are trusting them.

Recommendations:

- As we can observe a clear hike, we can have a good future in Brazil in terms of business but proceed with good pre-planning to gain the customer trust and belief.

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



Row	customer_state	avg_price	avg_freight	price_sum	freigthvalue_sum
1	SP	109.65	15.15	5202955.05	718723.07
2	PR	119.0	20.53	683083.76	117851.68
3	RS	120.34	21.74	750304.02	135522.74
4	MG	120.75	20.63	1585308.03	270853.46
5	ES	121.91	22.06	275037.31	49764.6
6	SC	124.65	21.47	520553.34	89660.26
7	RJ	125.12	20.96	1824092.67	305589.31
8	DF	125.77	21.04	302603.94	50625.5
9	GO	126.27	22.77	294591.95	53114.98
10	BA	134.6	26.36	511349.99	100156.68

```
select c.customer_state,

round(avg(price),2) as avg_price,
round(avg(freight_value),2) as avg_freight,
round(sum(price),2) as price_sum,
round(sum(freight_value),2) as freighvalue_sum
from `america_retailer_business.customers` c
join `america_retailer_business.orders` o
on c.customer_id = o.customer_id
join `america_retailer_business.order_items` ot
on o.order_id = ot.order_id
group by 1
order by 1
```

Insights:

- We can observe that different states have different freight values. Especially states like PB, RR, and CE have relatively higher mean freight values.
- States like PB, AL, AC, and PB have higher mean prices where customers are contributing more revenue to the company.
- States like SP, PR, RS, and MG have relatively lower mean prices.

Recommendations:

- We should try to reduce the freight values in states like PB, PR and CE by increasing partnership with local logistics and try to maintain inventory in multiple places.
- States where sum of prices are high, which shows there we can attract new customers and increase the market size and try to reduce the price of products and also should maintain the sufficient inventory.

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:

Row	ORDER_ID	time_to_delivery	diff_estimated_deliv
1	1950d777989f6a877539f5379...	30	-12
2	2c45c33d2f9cb8ff8b1c86cc28...	30	28
3	65d1e226dfaeb8cdc42f66542...	35	16
4	635c894d068ac37e6e03dc54e...	30	1
5	3b97562c3aee8bdedcb5c2e45...	32	0
6	68f47f50f04c4cb6774570cfde...	29	1
7	276e9ec344d3bf029ff83a161c...	43	-4
8	54e1a3c2b97fb0809da548a59...	40	-4
9	fd04fa4105ee8045f6a0139ca5...	37	-1
10	302bb8109d097a9fc6e9cefc5...	33	-5

3. Group data by state, take mean of freight_value, time_to_delivery, diff_estimated_delivery

Row	customer_state	mean_freight_value	time_to_delivery	diff_estimated_deliv
1	AC	40.07	20.33	20.01
2	AL	35.84	23.99	7.98
3	AM	33.21	25.96	18.98
4	AP	34.01	27.75	17.44
5	BA	26.36	18.77	10.12
6	CE	32.71	20.54	10.26
7	DF	21.04	12.5	11.27
8	ES	22.06	15.19	9.77
9	GO	22.77	14.95	11.37
10	MA	38.26	21.2	9.11

```
select customer_state,round(avg(freight_value),2) as mean_freight_value,
round(avg(TIMESTAMP_DIFF(TIMESTAMP(order_delivered_customer_date),
TIMESTAMP(order_purchase_timestamp), day)),2)
as time_to_delivery,
round(avg(TIMESTAMP_DIFF(TIMESTAMP(order_estimated_delivery_date),
TIMESTAMP(order_delivered_customer_date),day)),2)
as diff_estimated_delivery
from `america_retailer_business.orders` o
join `america_retailer_business.order_items` ot
on o.order_id = ot.order_id
join `america_retailer_business.customers` c
on o.customer_id = c.customer_id
group by 1
order by 1
```

Insights:

- Maximum time taken to deliver a product is 209 days which can impact our business due such delay in delivering the item which can lead to customer dissatisfaction.
- Maximum estimated time to deliver a product from the table is 155 days which is not a good sign to business and which can be improved.

Recommendations:

- We should try to reduce the number of days to deliver the product and try to update the customer why it is getting delay in delivering.
- Estimated delivery dates can also be reduced by following best inventory management practices and deals with local logistic businesses.

5.Top 5 states with highest/lowest average freight value - sort in desc/asc limit 5

States with highest average freight value

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

```
SELECT customer_state, ROUND(AVG(freight_value), 2) AS avg_freightvalue
FROM `america_retailer_business.customers` c
JOIN `america_retailer_business.orders` o ON c.customer_id = o.customer_id
JOIN `america_retailer_business.order_items` ot ON o.order_id = ot.order_id
GROUP BY 1
ORDER BY avg_freightvalue DESC
limit 5;
```

States with lowest average freight value

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

```
SELECT customer_state, ROUND(AVG(freight_value), 2) AS avg_freightvalue
FROM `america_retailer_business.customers` c
JOIN `america_retailer_business.orders` o ON c.customer_id = o.customer_id
JOIN `america_retailer_business.order_items` ot ON o.order_id = ot.order_id
GROUP BY 1
ORDER BY avg_freightvalue
limit 5;
```

Insights:

- Above table shows top 5 states which has highest average freight value which is not a good sign to the company and it is important thing where we should focus else, we might end in losses.

Recommendations:

- We should evaluate why average freight values of a state is high and adopt some techniques like local logistics delivery partners and maintaining good relationships with them.
- Planning to reduce freight values which can reduce cost to the company and maintain the economy of a business.

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

States with average highest time to delivery

Row	customer_state	time_to_delivery
1	RR	28.98
2	AP	26.73
3	AM	25.99
4	AL	24.04
5	PA	23.32

```
SELECT customer_state,

round(avg(TIMESTAMP_DIFF(TIMESTAMP(order_delivered_customer_date),
TIMESTAMP(order_purchase_timestamp), day)),2)
as time_to_delivery
FROM `america_retailer_business.customers` c
JOIN `america_retailer_business.orders` o ON c.customer_id = o.customer_id
GROUP BY 1
ORDER BY time_to_delivery DESC
limit 5;
```

States with lowest average time to delivery.

Row	customer_state	time_to_delivery
1	SP	8.3
2	PR	11.53
3	MG	11.54
4	DF	12.51
5	SC	14.48

```
SELECT customer_state,
round(avg(TIMESTAMP_DIFF(TIMESTAMP(order_delivered_customer_date),
TIMESTAMP(order_purchase_timestamp), day)),2)
as time_to_delivery
FROM `america_retailer_business.customers` c
JOIN `america_retailer_business.orders` o ON c.customer_id = o.customer_id
GROUP BY 1
ORDER BY time_to_delivery
limit 5;
```

Insights:

- Table shows top 5 states which has highest average time for delivery which leads to the customer dissatisfaction and sometimes even the loss of customers and because of the cost to deliver the product also increases which is again loss to the company.

Recommendations:

- Average time to delivery and freight values are directly proportional to each other which can clearly might be the reason for losses to company and also customer dissatisfaction.

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

Row	customer_state	diff_estimated_delivery
1	AL	7.95
2	MA	8.77
3	SE	9.17
4	ES	9.62
5	BA	9.93

States with lowest estimated delivery time difference

```
SELECT customer_state,
round(avg(TIMESTAMP_DIFF(TIMESTAMP(order_estimated_delivery_date),
TIMESTAMP(order_delivered_customer_date), day)),2)
as diff_estimated_delivery
FROM `america_retailer_business.customers` c
JOIN `america_retailer_business.orders` o ON c.customer_id = o.customer_id
GROUP BY 1
ORDER BY diff_estimated_delivery
limit 5;
```

States with highest estimated delivery time difference

Row	customer_state	diff_estimated_delive
1	AC	19.76
2	RO	19.13
3	AP	18.73
4	AM	18.61
5	RR	16.41

```
SELECT customer_state,

round(avg(TIMESTAMP_DIFF(TIMESTAMP(order_estimated_delivery_date),
TIMESTAMP(order_delivered_customer_date), day)),2)
as diff_estimated_delivery
FROM `america_retailer_business.customers` c
JOIN `america_retailer_business.orders` o ON c.customer_id = o.customer_id
GROUP BY 1
ORDER BY diff_estimated_delivery desc
limit 5;
```

6. Payment type analysis:

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

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

```
SELECT

extract(year from order_purchase_timestamp) as years
, extract(month from order_purchase_timestamp) as months
, payment_type,count(o.order_id) as number_of_orders,

FROM `america_retailer_business.payments` p
JOIN `america_retailer_business.orders` o
ON o.order_id = p.order_id
GROUP BY 1,2,3
order by 1,2
```

Insights:

- We can observe that, payments via credit cards are high in number.
- Second highest is UPI payments but they are consistent through out the time period which we have chosen.
- Debit card payments are less compared to other payment methods which can also be encouraged by us.

Recommendations:

- We can offer discounts to the customers like no cost EMI as possible as we can without effecting our businesses by tying partnership with the banks.
- Debit card transactions can also be increased by offering instant discounts to the customers.
- UPI based transaction are maintaining consistency which can be increased by making the process user friendly to use UPI and providing vouchers to the respective customers.

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

Row	payment_installment	number_of_orders
1	0	2
2	1	52546
3	2	12413
4	3	10461
5	4	7098
6	5	5239
7	6	3920
8	7	1626
9	8	4268
10	9	644

```
SELECT
    extract(year from order_purchase_timestamp) as years
    , extract(month from order_purchase_timestamp) as months
    , payment_type,count(o.order_id) as number_of_orders,

    FROM `america_retailer_business.payments` p
    JOIN `america_retailer_business.orders` o
    ON o.order_id = p.order_id
    GROUP BY 1,2,3
    order by 1,2
```

Insights:

- One-time payment is the most preferable by the customers which occupies almost half of the count of orders.
- Instalments up to 3 are considerable in amount of orders.

Recommendations:

- We can provide offers and vouchers for multiple instalment payments to encourage the customers to pay in instalments.
- By understanding the preferences of customers we can provide discounts based on their comfort.