Retail Store CS

**To quickly search columns in BigQuery use Dataplex Catalog:
https://console.cloud.google.com/dataplex/dp-search?project=feb24scale**

Q1.1 Data type of all columns in the "customers" table.

In BigQuery the db naming convention is: **feb24scale** is project name, **retial_cs** is DB name.

Ans:
```
SELECT
  column_name,
  data_type
FROM
  `feb24scale.retial_cs.INFORMATION_SCHEMA.COLUMNS`
WHERE
  table_name = 'customers';
```

Screenshot:

Field name	Type	Mode
customer_id	STRING	NULLABLE
customer_unique_id	STRING	NULLABLE
customer_zip_code_prefix	INTEGER	NULLABLE
customer_city	STRING	NULLABLE
customer_state	STRING	NULLABLE

Insights: NA

Recommendation: NA

Q1.2 Get the time range between which the orders were placed.

Ans:
```sql
SELECT
  MIN(order_purchase_timestamp) AS `min_purchase_time`,
  MAX(order_purchase_timestamp) AS `max_purchase_time`
FROM
  `retial_cs.orders`
```

Screenshot:

Row	min_purchase_time ▼	max_purchase_time ▼
1	2016-09-04 21:15:19 UTC	2018-10-17 17:30:18 UTC

Insights:
1. For the current data set the orders are placed from 4th September, 2016 from 9:15 PM to 17th November 2018 until 5:30 PM.

Recommendation:
1. Its better to look at peak or min sales times to get better insights.

Q1.3 Count the Cities & States of customers who ordered during the given period.

Ans:
```sql
SELECT
  COUNT(DISTINCT customer_city) AS `city_count`,
  COUNT(DISTINCT customer_state) AS `state_count`
FROM
  `retial_cs.customers`
```

Screenshot:

Row	city_count ▼	state_count ▼
1	4119	27

Insights:
1. There are 27 states in total and it is not surprising that all 27 states made an order in that period. However there are a total of 8011 cities and approx half of them only seems to have made an order. So its safe to conclude we do not have any presence in those cities.
2. I cannot conclude on customer engagement in this but since the 4119 cities did place at least one order it is likely that they will place again so we can consider starting loyalty programs in these cities if its not already in place.

```
SELECT
  COUNT(DISTINCT geolocation_state) AS `state_count`,
  COUNT(DISTINCT geolocation_city) AS `city_count`
FROM
  `cs1.geolocation`
```

Recommendation:
1. We can try setting up new stores and expand our reach in the cities which did not make any orders through targeted marketing.

Q2.1 Is there a growing trend in the no. of orders placed over the past years?

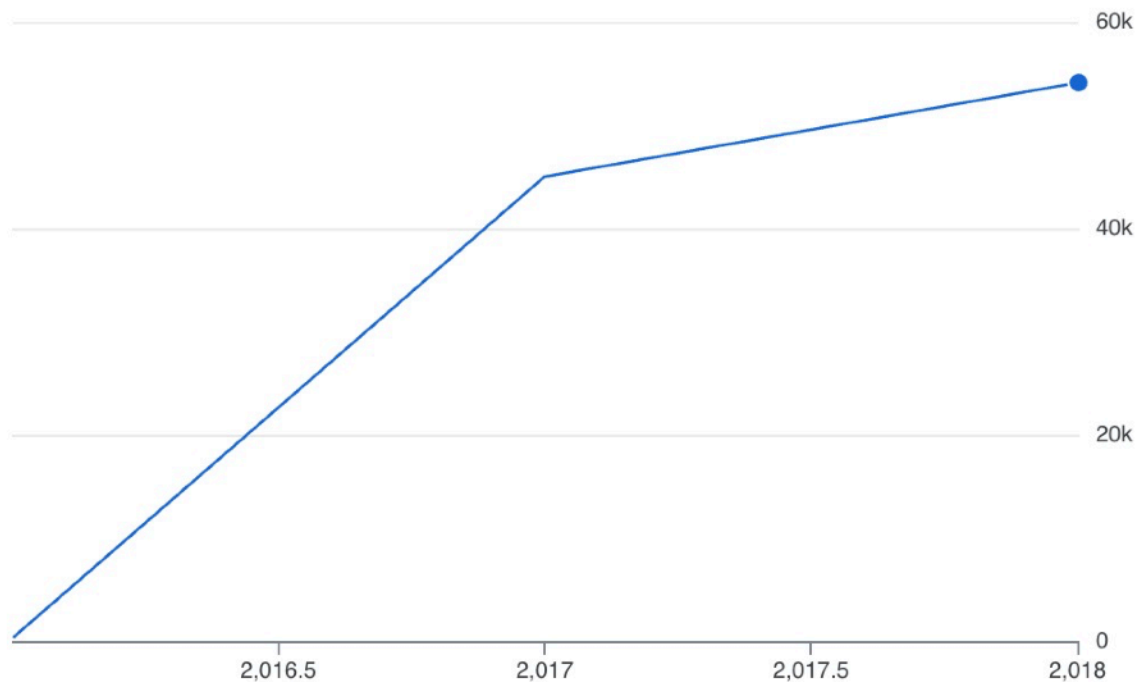Trend is observed by looking at the flow over a period of time.

Ans:
```
SELECT
  EXTRACT(YEAR FROM order_purchase_timestamp) AS `year`,
  COUNT(*) AS `total_orders_per_year`
FROM
  `cs1.orders`
GROUP BY
  year
ORDER BY
  year
```

Screenshot:

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

total_orders_per_year by year

Insights:
1. Yes there is a growing trend in the number of orders placed. A great spike from in sales from 2016 to 2017. Probably due to a marketing campaign or due to new store opening in new locations.
2. From the given data we can say there is a growth in the orders placed from 2017 to 2018. However the data is insufficient to predict continuous growth in the future. Sample size is too small.

Recommendation:
1. We should emulate the growth strategy of 2016-2017 to see if the spike is reproducible assuming its not a new store opening.
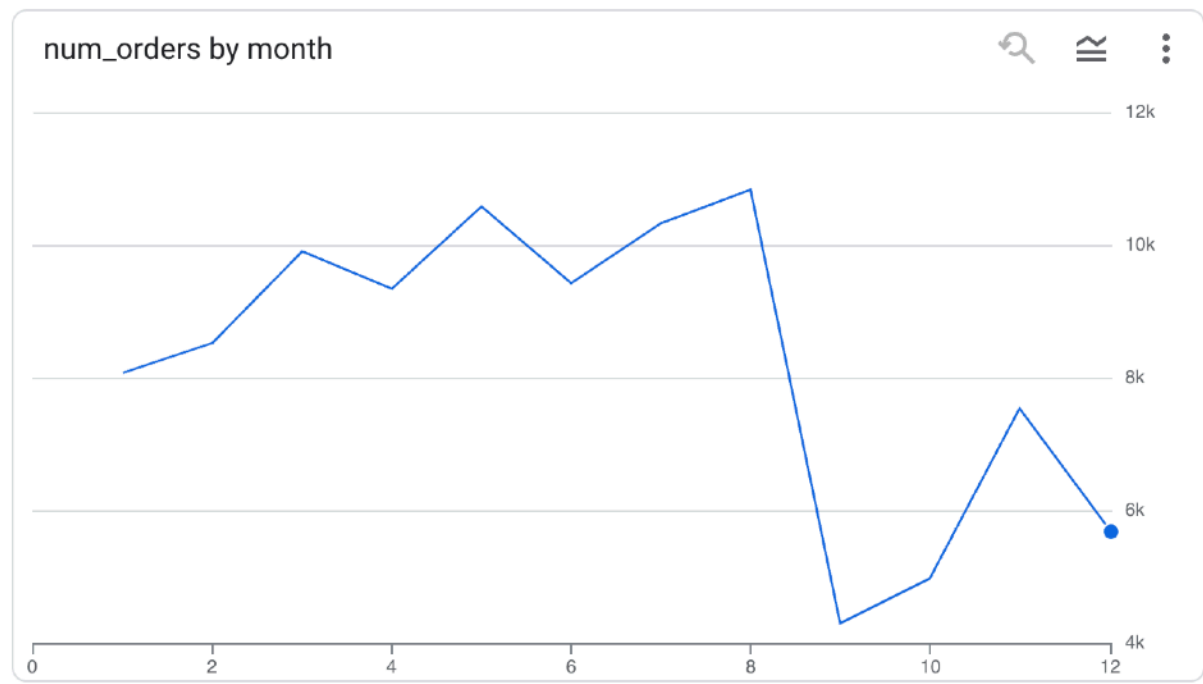
_____

_____

Q2.2 Can we see some kind of monthly seasonality in terms of the no. of orders being placed?

Ans:
```sql
SELECT
  EXTRACT(MONTH FROM order_purchase_timestamp) AS `month`,
  COUNT(*) AS `num_orders`
FROM
  `retial_cs.orders`
GROUP BY
  month
ORDER BY
  month ASC;
```

Screenshot:

Row	month	num_orders
1	1	8069
2	2	8508
3	3	9893
4	4	9343
5	5	10573
6	6	9412
7	7	10318
8	8	10843
9	9	4305
10	10	4959

num_orders by month

Insights:
1. Yes there is seasonality. Especially there is a drastic drop in sales during September to December months probably due to festival season or winter break.

Recommendation:
1. Adjust the inventory during months 9 to 12 by either performing flash sale or bundled/up sale to exhaust the inventory. Or reduce the store working hours to avoid unnecessary store maintenance costs. Or stock up specific products that are related to the festival to increase sales.

Q2.3 During what time of the day, do the Brazilian customers mostly place their orders? (Dawn, Morning, Afternoon or Night)

Ans:
```sql
SELECT
  CASE
    WHEN EXTRACT(HOUR FROM order_purchase_timestamp) BETWEEN 0
AND 6 THEN 'Dawn'
    WHEN EXTRACT(HOUR FROM order_purchase_timestamp) BETWEEN 7
AND 12 THEN 'Morning'
    WHEN EXTRACT(HOUR FROM order_purchase_timestamp) BETWEEN
13 AND 18 THEN 'Afternoon'
    ELSE 'Night'
  END AS time_of_day,
  COUNT(*) AS num_orders
```

```
FROM `retial_cs.orders`
GROUP BY 1
ORDER BY 2 DESC;
```

Screenshot:

Row	time_of_day ▼	num_orders ▼
1	Afternoon	38135
2	Night	28331
3	Morning	27733
4	Dawn	5242

num_orders by time_of_day
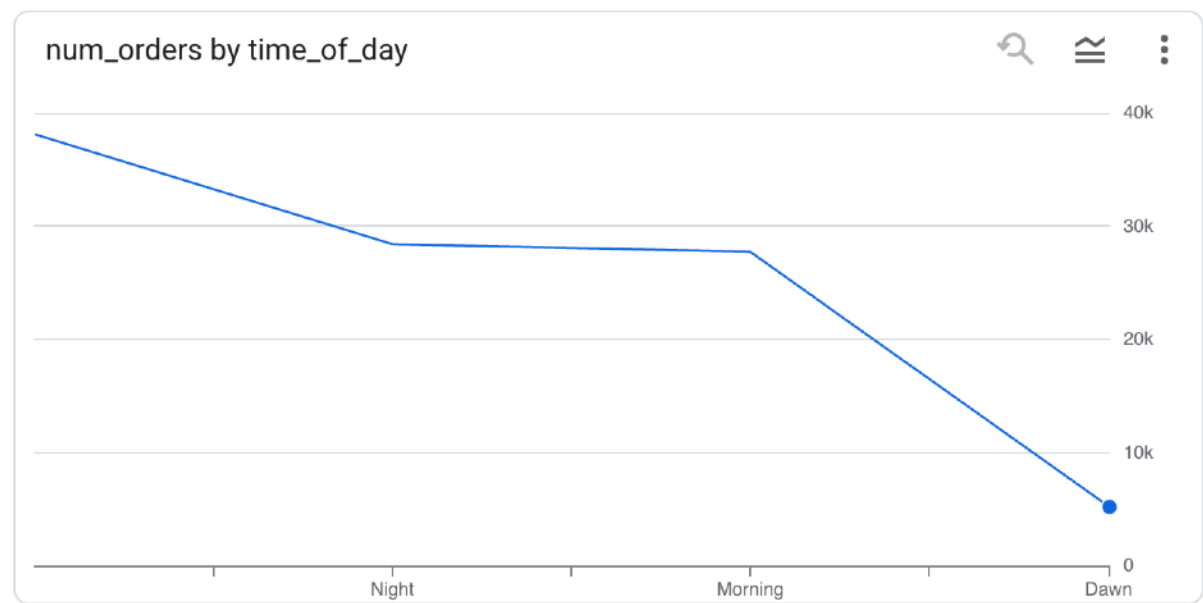
Insights:
1. Brazilian customers mostly place their orders in the afternoon.

Recommendation:
1. Adjust store timings, increase staff capacity and stock up more products during afternoons to address more customers in Brazil location.

Q3.1 Get the month on month no. of orders placed in each state.

Ans:
```
SELECT
  c.customer_state,
```

```
    EXTRACT(MONTH FROM order_purchase_timestamp) AS month,
    COUNT(*) AS num_orders
FROM `retial_cs.orders` AS o
JOIN `retial_cs.customers` AS c ON o.customer_id =
c.customer_id
GROUP BY 1, 2
ORDER BY 1, 2;
```
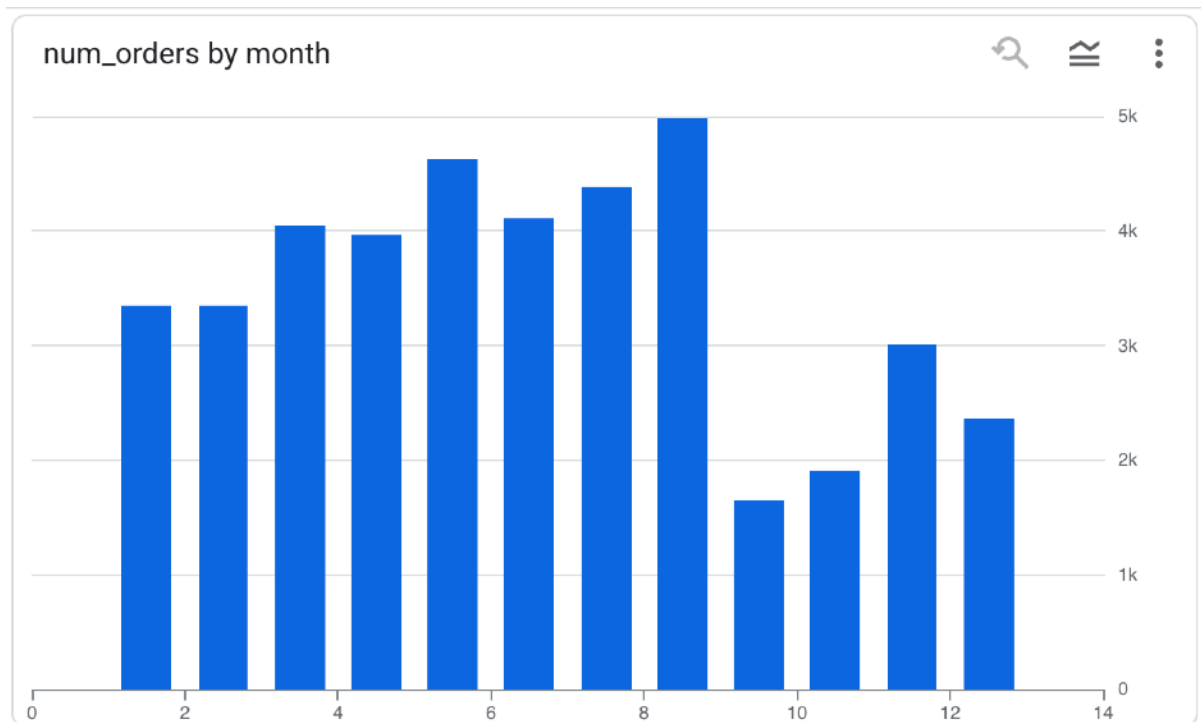
Screenshot:

Row	customer_state ▾	month ▾	num_orders ▾
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

num_orders by customer_state

num_orders by month
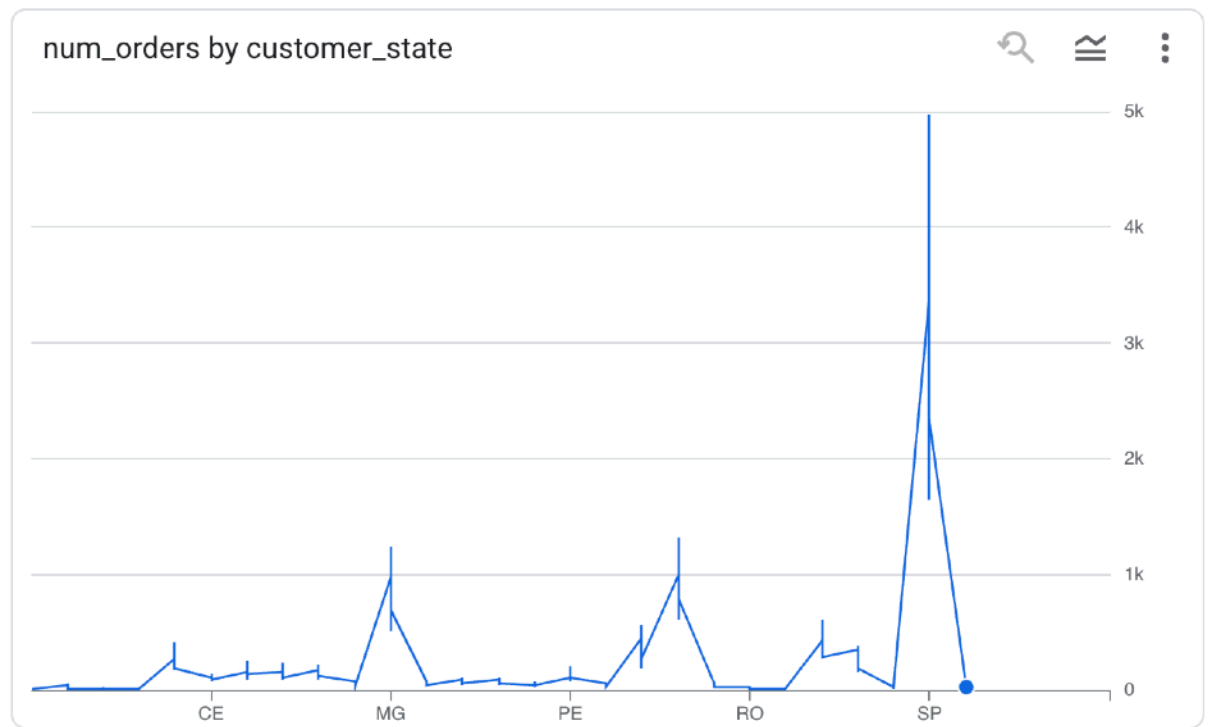
Insights:
1. It appears that some states have low activity in November and December months while others dont. Unsurprisingly. So we can conclude that the the states with low sales during 9th and 10th months are either in winter break or some kind of festival season.

Recommendation:
1. We can adjust stock based on season in every location. Follow Q2.3 recommendations.

Q3.2 How are the customers distributed across all the states?

Ans:
```
SELECT
  customer_state,
  COUNT(DISTINCT customer_unique_id) AS num_customers
FROM `retial_cs.customers`
GROUP BY 1
ORDER BY 2 DESC;
```

Screenshot:

Row	customer_state	num_customers
1	SP	40302
2	RJ	12384
3	MG	11259
4	RS	5277
5	PR	4882
6	SC	3534
7	BA	3277
8	DF	2075
9	ES	1964
10	GO	1952

num_customers by customer_state
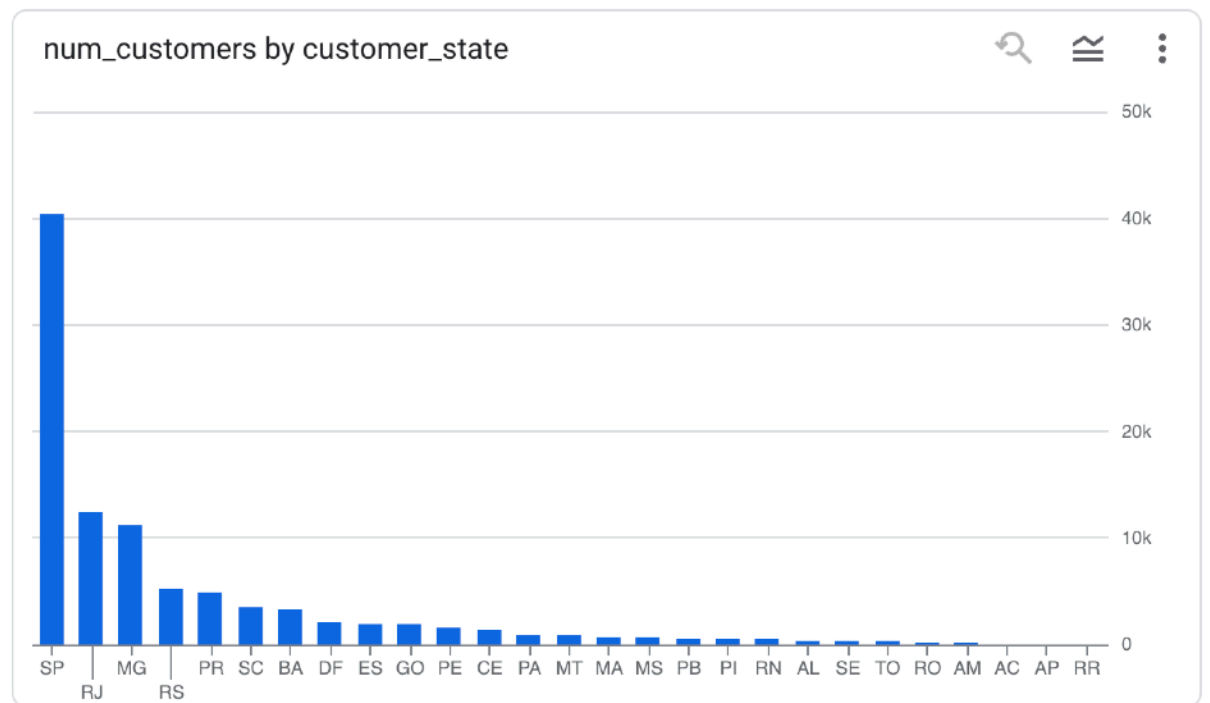
Insights:
1. State SP has highest customers with 41746 while state RR has lowest with 46.

Recommendation:

1. We should focus more on SP state to drive more sales and set up loyalty programs to retain existing customers better. We should promote more in RR and other lesser states. Either our promotions are not effective or the price may be too high for this area.

Q4.1 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.

Algo:
1. Get cost of orders in Jan to Aug 2017 and add all of them.
2. Get cost of orders in Jan to Aug 2018 and add all of them.
3. Subtract both of them to get the increase.
4. To get % increase use this formula: 100 * ((final - initial) / initial)

Ans:
```
WITH cost_2017 AS (
  SELECT SUM(payment_value) AS total_2017
  FROM `retial_cs.payments` AS p
  JOIN `retial_cs.orders` AS o ON p.order_id = o.order_id
  WHERE o.order_purchase_timestamp BETWEEN '2017-01-01' AND
'2017-08-31'
),
cost_2018 AS (
  SELECT SUM(payment_value) AS total_2018
  FROM `retial_cs.payments` AS p
  JOIN `retial_cs.orders` AS o ON p.order_id = o.order_id
  WHERE o.order_purchase_timestamp BETWEEN '2018-01-01' AND
'2018-08-31'
)
SELECT
  ROUND(((total_2018 - total_2017) / total_2017) * 100, 2) AS
percent_increase
FROM cost_2017, cost_2018;
```

Screenshot:

Row	percent_increase
1	138.53

Insights:
1. 5025712 increase in amount or 137 % increase in payments from 2017 to 2018.
2. January to August is the time when we had high sales as per the given data. Stretched over time we can get an insight in the customer spending behaviour.
3. It appears that there is demand for our products and customers are willing to either pay more or we can sell more.

Recommendation: NA

Q4.2 Calculate the Total & Average value of order price for each state.

Ans:
```sql
SELECT
  c.customer_state AS state,
  ROUND(SUM(oi.price), 2) AS total_price,
  ROUND(AVG(oi.price), 2) AS avg_price
FROM
  (`retial_cs.order_items` AS oi JOIN `retial_cs.orders` AS o
ON
  oi.order_id = o.order_id) JOIN `retial_cs.customers` AS c ON
  c.customer_id = o.customer_id
GROUP BY 1
ORDER BY 1
```

Screenshot:

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

total_price, avg_price by state

Insights:
1. The highest revenue is from SP state at 5202955.05 while least sale happened in RR state at 7829.43. However the highest average price is from PB state at 191.48 and least from SP at 109.65. So it appears there are much more wealthy customers or products are in demand in PB and SP states particularly in PB state with high average order price.

Recommendation:
1. Investing more in PB might give us higher returns in the future. We should prepare premium plans for these areas and perform a test run to see if it drives more revenue.

Q4.3 Calculate the Total & Average value of order freight for each state.

Ans:
```sql
SELECT
  c.customer_state AS `state`,
  ROUND(SUM(oi.freight_value), 2) AS `total_freight`,
  ROUND(AVG(oi.freight_value), 2) AS `avg_freight`
FROM
  (`retial_cs.order_items` AS `oi` JOIN `retial_cs.orders` AS `o` ON
  oi.order_id = o.order_id) JOIN `retial_cs.customers` AS `c`
ON
```
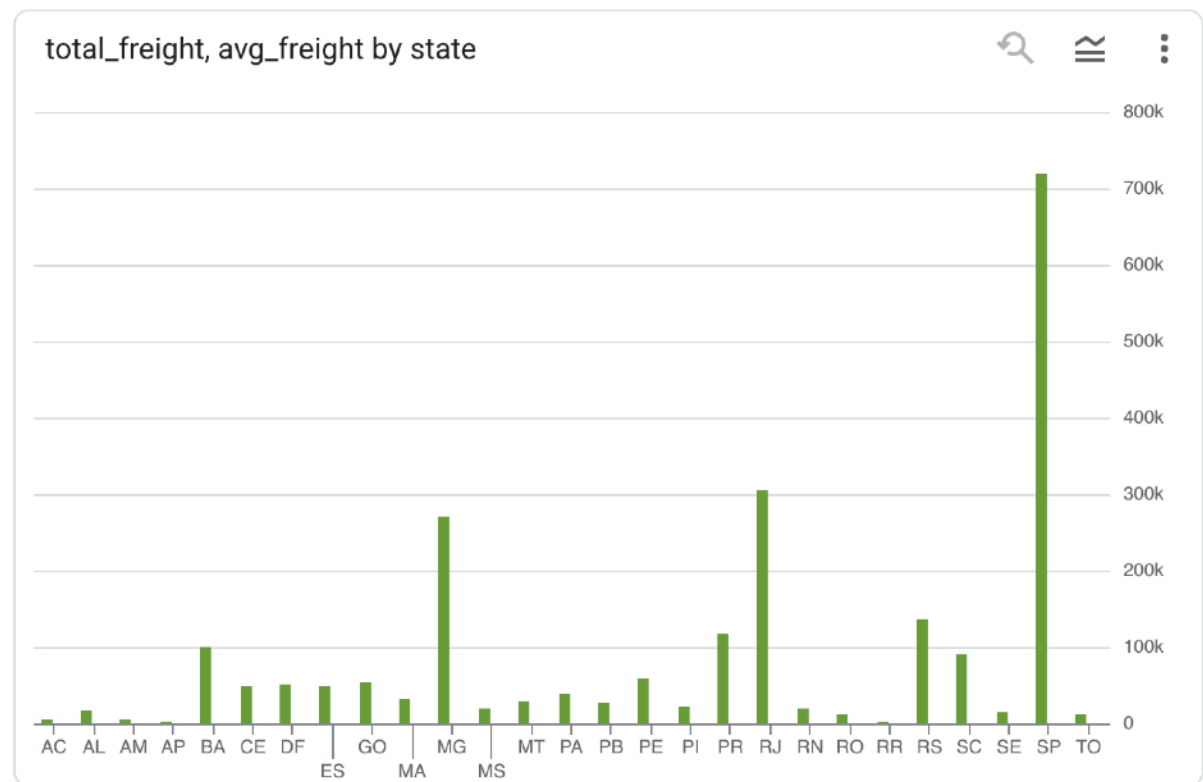
```
    c.customer_id = o.customer_id
GROUP BY 1
ORDER BY 1
```

Screenshot:

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

total_freight, avg_freight by state

Insights:
1. SP state has highest freight value at 718723.07 while RR seems to have lowest at 2235.19. Avg wise RR has highest at 42.98 while SP has lowest at 15.15.

Recommendation:
1. Find ways to reduce packaging size in area SP to reduce freight costs.
2. Find alternate transportation or routes.

Q5.1 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:

Ans:
```sql
SELECT
  order_id,
  ABS(DATE_DIFF(order_purchase_timestamp,
order_delivered_customer_date, DAY)) AS `time_to_deliver`,
  ABS(DATE_DIFF(order_estimated_delivery_date,
order_delivered_customer_date, DAY)) AS
`diff_estimated_delivery`
FROM
  `retial_cs.orders`
WHERE
  order_delivered_customer_date IS NOT NULL
ORDER BY 2 DESC;
```

Screenshot:

Row	order_id	time_to_deliver	diff_estimated_delivery
1	ca07593549f1816d26a572e06d...	209	181
2	1b3190b2dfa9d789e1f14c05b6...	208	188
3	440d0d17af552815d15a9e41a...	195	165
4	285ab9426d6982034523a855f...	194	166
5	0f4519c5f1c541ddec9f21b3bd...	194	161
6	2fb597c2f772eca01b1f5c561bf...	194	155
7	47b40429ed8cce3aee9199792...	191	175
8	2fe324febf907e3ea3f2aa96508...	189	167
9	2d7561026d542c8dbd8f0daead...	188	159
10	c27815f7e3dd0b926b58552628...	187	162

Insights:
1. There seems to be a gap of min 20 days/hours in the expectations vs reality of delivery.
2. Inefficient route planning or poor algorithms or poor overall planning.

Recommendation:
1. Improve path finding algorithms to better plan the route if being used to ensure timely order fulfilment.
2. Measure time at every checkpoint to identify delivery delays to optimise supply chain processes.
3. Find ways to reduce packaging size to reduce freight costs.
4. Find alternate transportation or routes.

Q5.2 Find out the top 5 states with the highest & lowest average freight value.

Ans:
```sql
-- Highest Freight
SELECT
  customer_state,
  ROUND(AVG(freight_value), 2) AS `avg_freight`
FROM `retial_cs.order_items` AS `oi`
JOIN `retial_cs.orders` AS `o` ON oi.order_id = o.order_id
JOIN `retial_cs.customers` AS `c` ON o.customer_id =
c.customer_id
GROUP BY 1
ORDER BY 2 DESC
LIMIT 5;

-- Lowest Freight (Replace DESC with ASC)
```

Screenshot:

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

avg_freight by customer_state
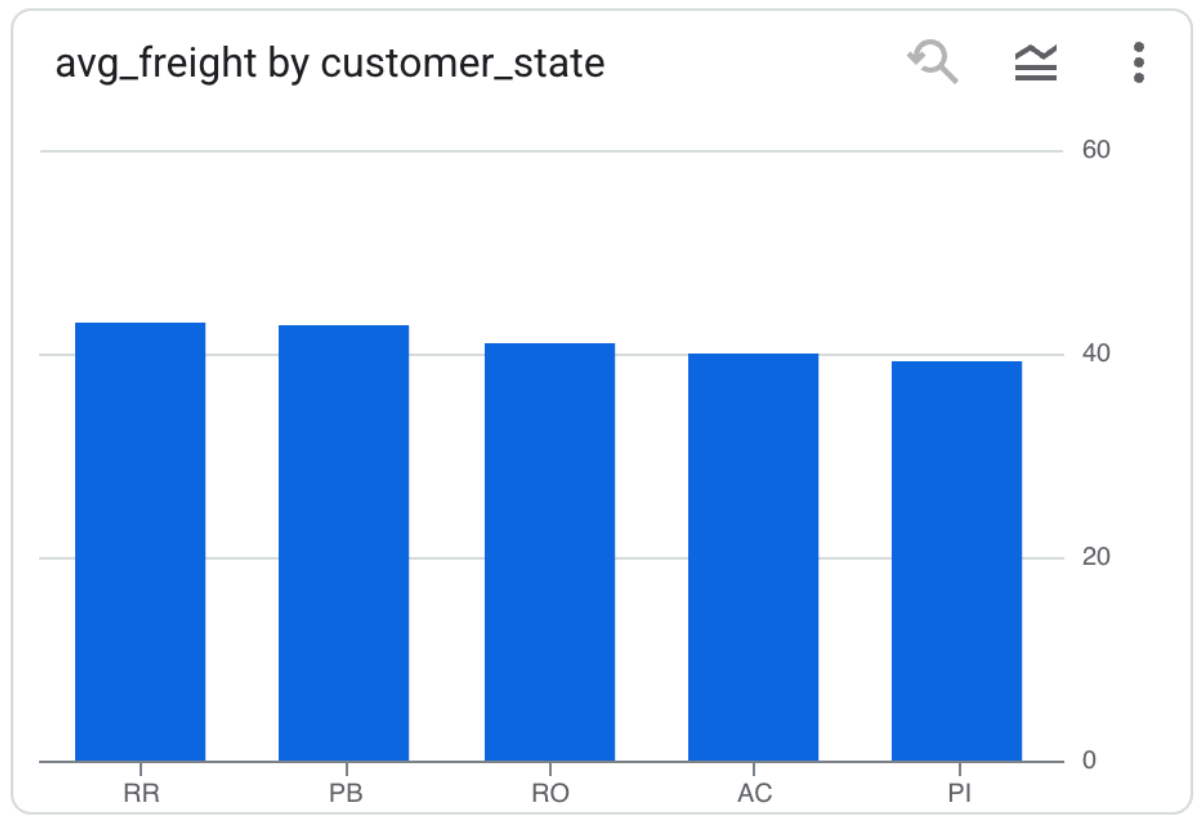
Insights:
1. RR has highest freight value at 42.98 while SP has lowest at 15.15.

Recommendation:
1. Optimise packaging accordingly to increase the quantity shipped.
2. Negotiate for lower freight rates.
3. Improve resource allocation.
4. Follow recommendations mentioned in Q5.1

**Q5.3 Find out the top 5 states with the highest & lowest average delivery time.**

Ans:
```sql
SELECT
  customer_state,
  ROUND(AVG(DATE_DIFF(o.order_estimated_delivery_date,
o.order_delivered_customer_date, DAY)), 2) AS `avg_diff`
FROM `retial_cs.orders` AS `o`
JOIN `retial_cs.customers` AS `c`
ON o.customer_id = c.customer_id
WHERE order_delivered_customer_date IS NOT NULL
GROUP BY 1
ORDER BY 2 DESC;
```

Screenshot:

1	AL	7.95
2	MA	8.77
3	SE	9.17
4	ES	9.62
5	BA	9.93
6	CE	9.96
7	SP	10.14
8	MS	10.17
9	PI	10.47
10	SC	10.61

avg_diff by customer_state
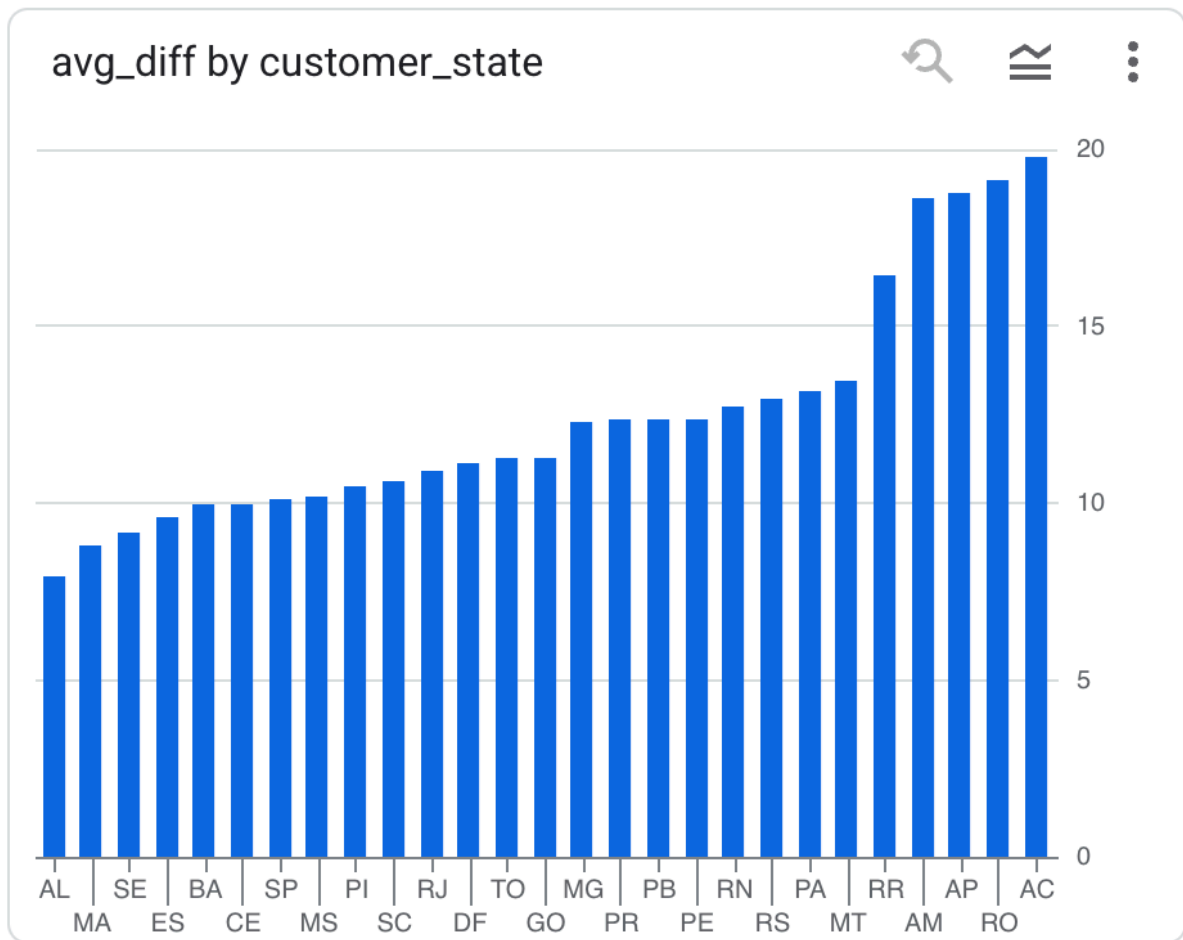
Insights:
1. AL state has highest delivery time while AC has lowest. There could be many reasons for this. Dense population, poor transportation, product demand, state laws, etc.
2. Lower the difference, higher the delivery time.

Recommendation:
1. Scout the area to get more info and optimise departments that are taking more time.
2. Delivery time and freight rates can sometimes have a correlation. Follow recommendations in Q5.2

Q5.4 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.

Ans:

Insights:
1. State AC is faster by 20.13 days from its estimated delivery date 40.77 days. ~1/2 the time.
2. Faster deliveries improve customer satisfaction.
3. Faster deliveries also shows a streamlined delivery process. The process can be used as a template to emulate the speed in other areas.

Recommendation:
1. Increase stock in AM state to make more deliveries.

Q6.1 Find the month on month no. of orders placed using different payment types.

Ans:
```sql
SELECT
  EXTRACT(MONTH FROM order_purchase_timestamp) AS `month`,
  payment_type,
  COUNT(DISTINCT o.order_id) AS `num_orders`
FROM `retial_cs.orders` AS `o`
JOIN `retial_cs.payments` AS `p`
ON o.order_id = p.order_id
GROUP BY 1, 2
ORDER BY 1, 2;
```

Screenshot:

Row	month	payment_type	num_orders
1	1	UPI	1715
2	1	credit_card	6093
3	1	debit_card	118
4	1	voucher	337
5	2	UPI	1723
6	2	credit_card	6582
7	2	debit_card	82
8	2	voucher	288
9	3	UPI	1942
10	3	credit_card	7682

num_orders by month
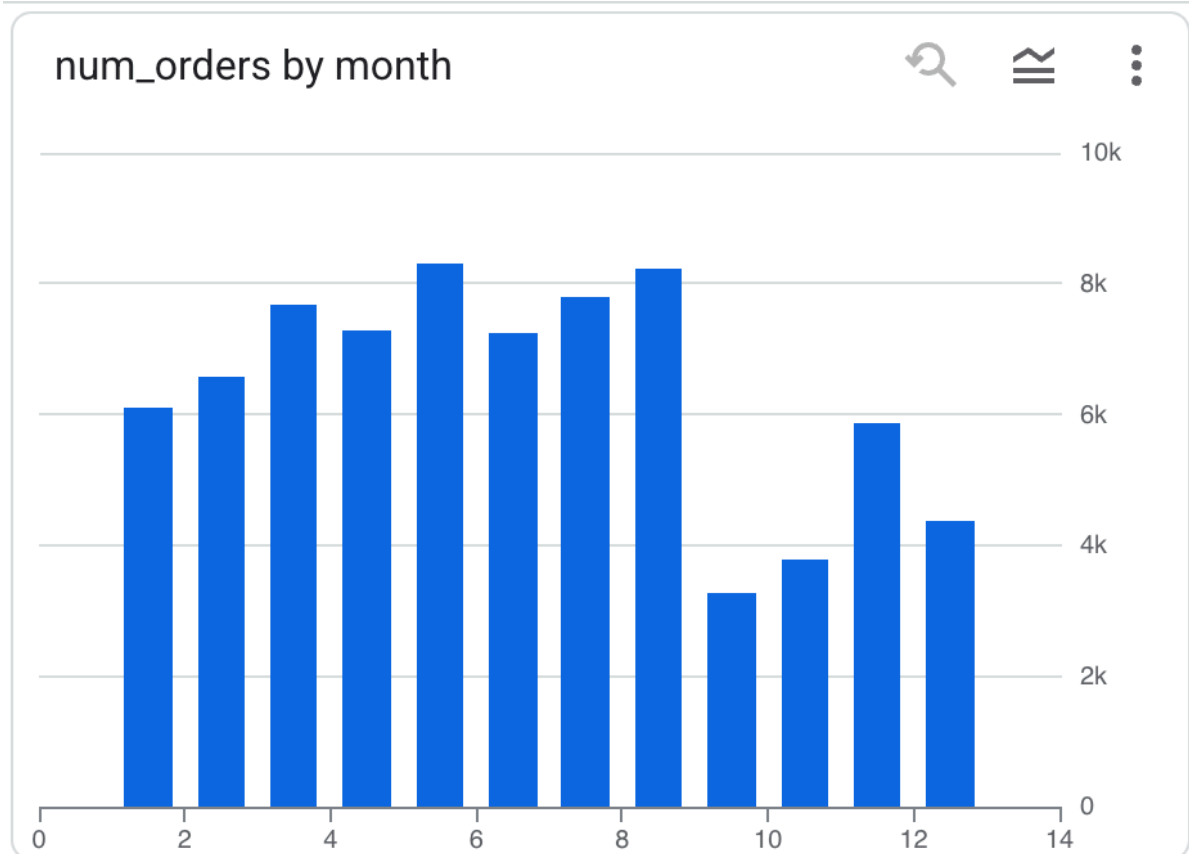
Insights:
1. Credit card is clearly the most popular choice followed by UPI. Debit card being the least preferred.
2. We can observe seasonality in the usage of payment types.

Recommendation:
1. We should team up with banks to give low interest credit to those who signup with our scheme. Provide cash backs for using their credit card.

_____

_____

Q6.2 Find the no. of orders placed on the basis of the payment installments that have been paid.

Ans:
```sql
SELECT
  payment_installments,
  COUNT(DISTINCT order_id) AS `num_orders`
FROM `retial_cs.payments`
GROUP BY 1
ORDER BY 1;
```

Screenshot:

Row	payment_installm...	num_orders
1	0	2
2	1	49060
3	2	12389
4	3	10443
5	4	7088
6	5	5234
7	6	3916
8	7	1623
9	8	4253
10	9	644

num_orders by payment_installments

Insights:
1. Highest payment installement is 1

1. Focus on High-Growth States: SP and RJ show the highest order volumes. Invest in logistics here.
2. Optimize Delivery Times: States like RR and AP have slow deliveries; improve carrier partnerships.
3. Leverage Payment Trends: Credit card usage dominates; offer installment promotions.
4. Seasonal Campaigns: Boost marketing during November (holiday season) and address January slumps.
5. Freight Cost Reduction: Negotiate lower freight rates in high-cost states (e.g., RR).