Business Case Study
TARGET SQL

Target, a globally recognized brand and a prominent retail force in the United States, solidifies its status as a preferred shopping destination by delivering exceptional value, inspiration, innovation, and an unmatched guest experience. This business case hones in on Target's operations in Brazil, providing in-depth insights into 100,000 orders spanning 2016 to 2018. The dataset offers a comprehensive perspective on crucial dimensions, encompassing order status, pricing, payment and freight performance, customer locations, product attributes, and customer reviews.

The analysis of this extensive dataset unveils valuable insights into Target's operations in the Brazilian market. It serves as a key to understanding various facets of the business, ranging from order processing and pricing strategies to the efficiency of payment and shipping processes. Additionally, it delves into the demographics of Target's customer base, the characteristics of the products offered, and the levels of customer satisfaction. By examining these dimensions, businesses can glean actionable intelligence, enabling them to refine strategies, enhance operational efficiency, and cater more precisely to the diverse needs and preferences of their Brazilian customer base.

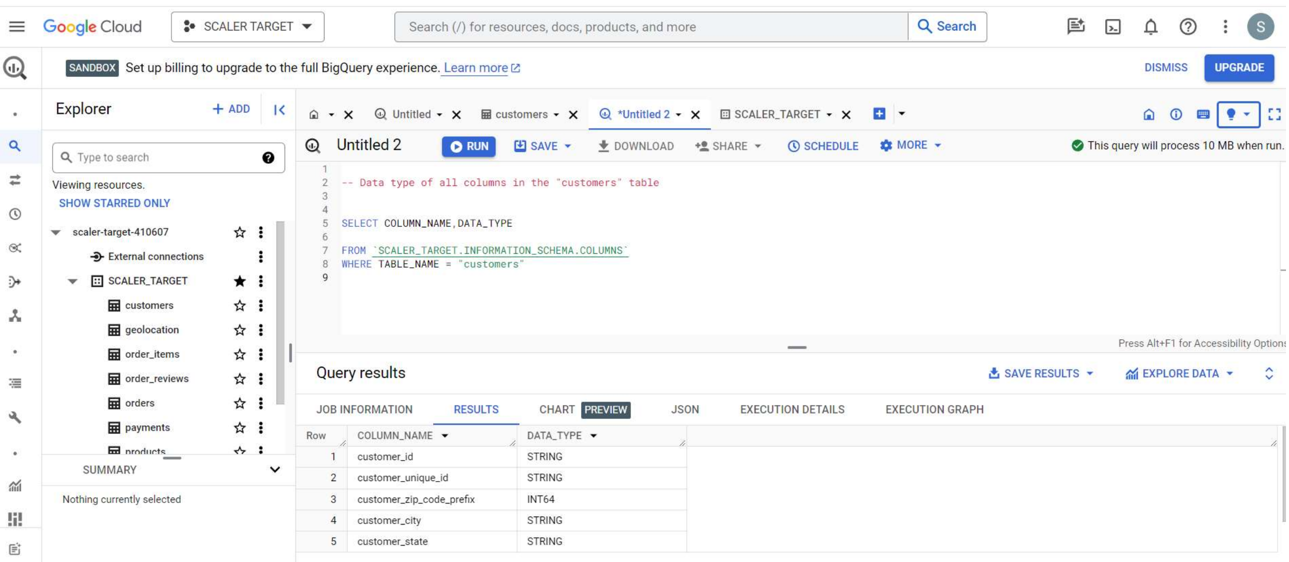
Task 1: Import the dataset and do usual exploratory analysis steps like checking the structure & characteristics of the dataset :

#Q1 - Data type of all columns in the "customers" table

Query:

```sql
SELECT COLUMN_NAME,DATA_TYPE

FROM `SCALER_TARGET.INFORMATION_SCHEMA.COLUMNS`
WHERE TABLE_NAME = "customers"
```

Output:



Insight & Recommendations

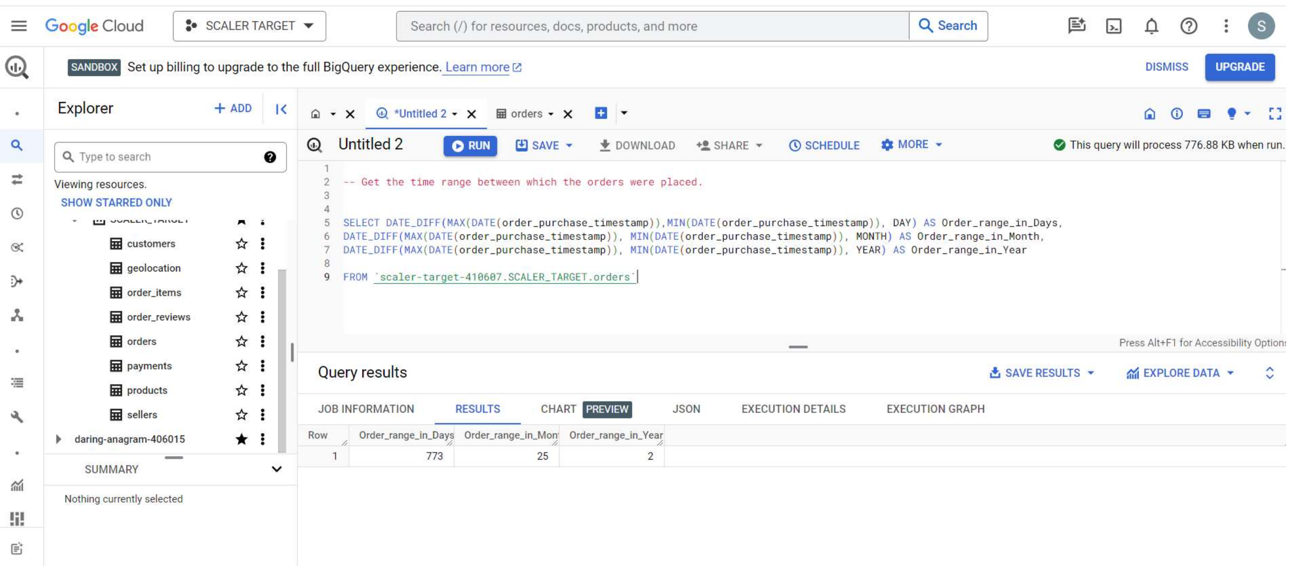
This columns are all saved as strings (VARCHAR), with the exception of the customer_zip_code_prefix column. Which is saved as Integer. This suggests that the majority of the data may be textual, and that any values pertaining to dates or numbers are most likely preserved as strings.

#Q2 - Get the time range between which the orders were placed.

Query:

```
SELECT
DATE_DIFF(MAX(DATE(order_purchase_timestamp)),MIN(DATE(ord
er_purchase_timestamp)), DAY) AS Order_range_in_Days,
DATE_DIFF(MAX(DATE(order_purchase_timestamp)),
MIN(DATE(order_purchase_timestamp)), MONTH) AS
Order_range_in_Month,
DATE_DIFF(MAX(DATE(order_purchase_timestamp)),
MIN(DATE(order_purchase_timestamp)), YEAR) AS
Order_range_in_Year
FROM `scaler-target-410607.SCALER_TARGET.orders`
```

Output:



Insight & Recommendations

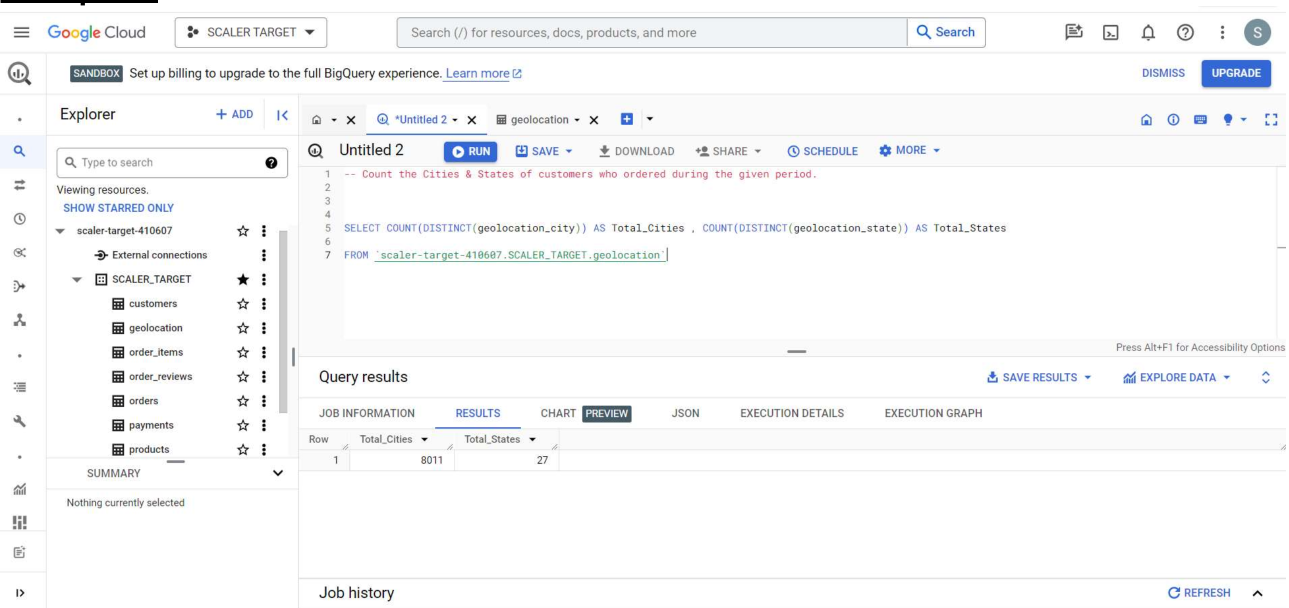
In this instance, the orders were placed for a **duration** of **2 years or 25 months or 773 days.**
The time range of the orders might be useful in analysing trends, market volatility, and overall order patterns along with products of interest over a given period of time.

#Q3 - Count the Cities & States of customers who ordered during the given period.

Query:

```
SELECT COUNT(DISTINCT(geolocation_city)) AS Total_Cities,
COUNT(DISTINCT(geolocation_state)) AS Total_States

FROM `scaler-target-410607.SCALER_TARGET.geolocation`
```

Output:



Insight & Recommendations

The dataset provides insights into the geographic distribution of customers with a total of **8011 different cities** throughout **27 different states**. Information on variety, concentration, and regional presence can be found by analysing the distribution of cities and states. This data facilitates the identification of hotspots and evaluates the degree of a company's national or worldwide reach.

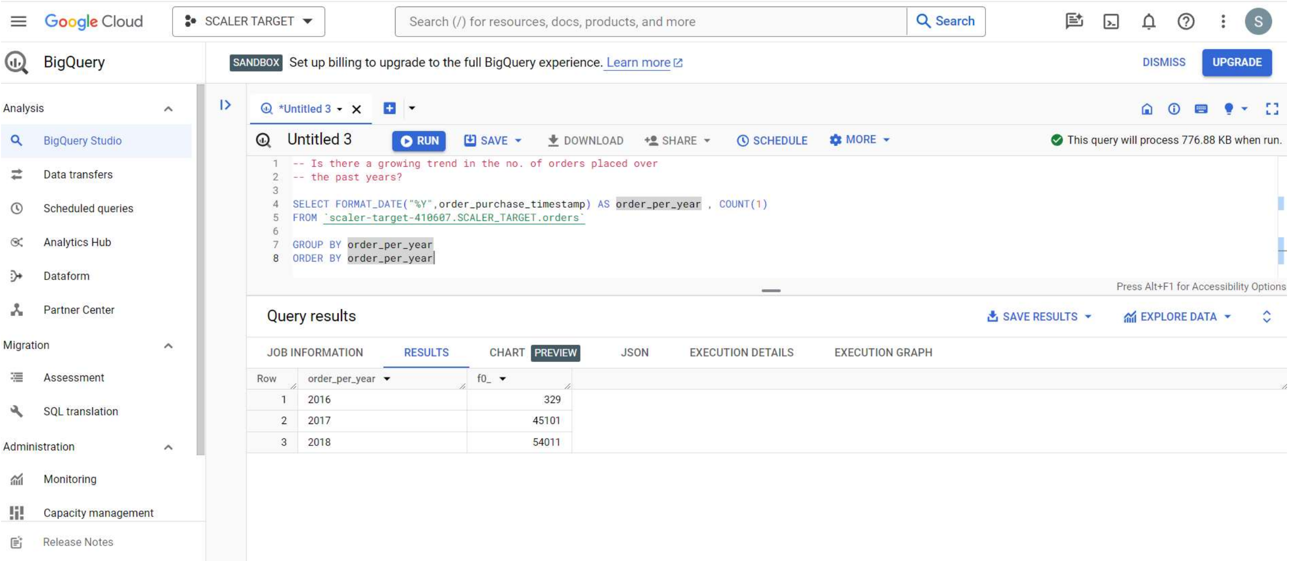
Task 2: In-Depth Exploration:

#Q1 – Is there a growing trend in the no. Of orders placed over the past years ?

Query:

```
SELECT FORMAT_DATE("%Y",order_purchase_timestamp) AS
order_per_year , COUNT(1)
FROM `scaler-target-410607.SCALER_TARGET.orders`

GROUP BY order_per_year
ORDER BY order_per_year
```

Output:



Row	order_per_year	f0_
1	2016	329
2	2017	45101
3	2018	54011

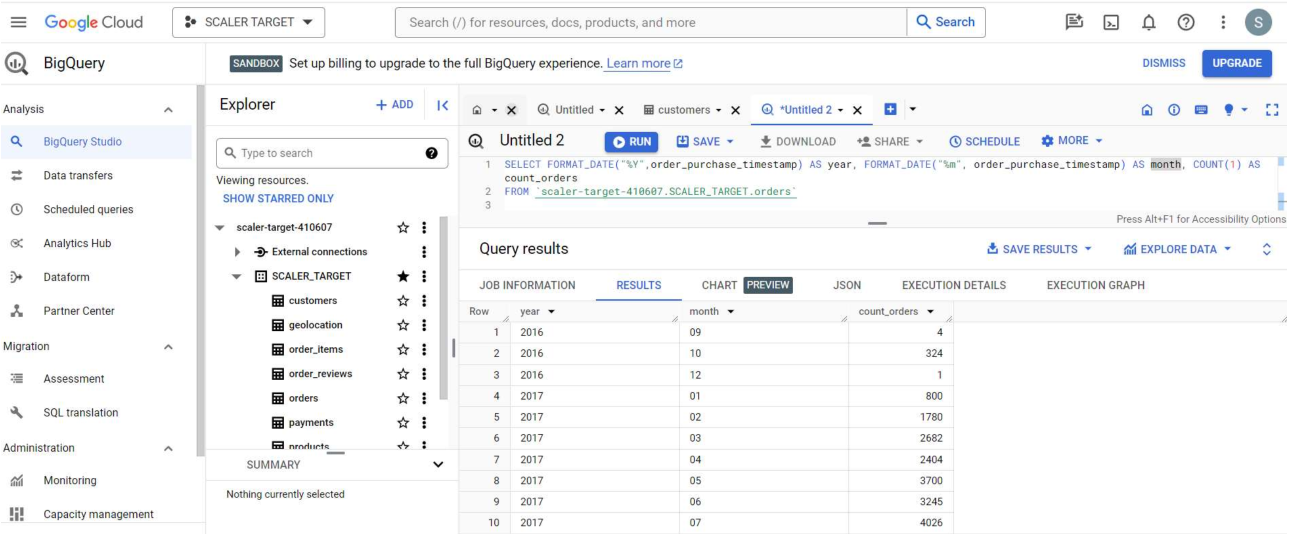
Insight & Recommendations A positive trend may be seen in the order volume, which has been steadily rising in recent years. A positive trend is indicated by consistent increase from year on year.

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

Query:

```
SELECT FORMAT_DATE("%Y",order_purchase_timestamp) AS year,
FORMAT_DATE("%m", order_purchase_timestamp) AS month,
COUNT(1) AS count_orders
FROM `scaler-target-410607.SCALER_TARGET.orders`

GROUP BY year,month
ORDER BY year,month
```

Output:



Insight & Recommendations

We observe a **seasonal pattern** in November 2017, the month of Black Friday, and a significant spike in orders. New Year's celebrations in January 2017 and January 2018 are also showing growing trends, and some may have placed pre-orders for the Carnival in February.

Comprehending monthly seasonality can be beneficial for consumer behaviour, marketing **strategies**, and **operational planning**. It can support more effective resource allocation, inventory management optimisation, peak time identification, and promotional activity planning.

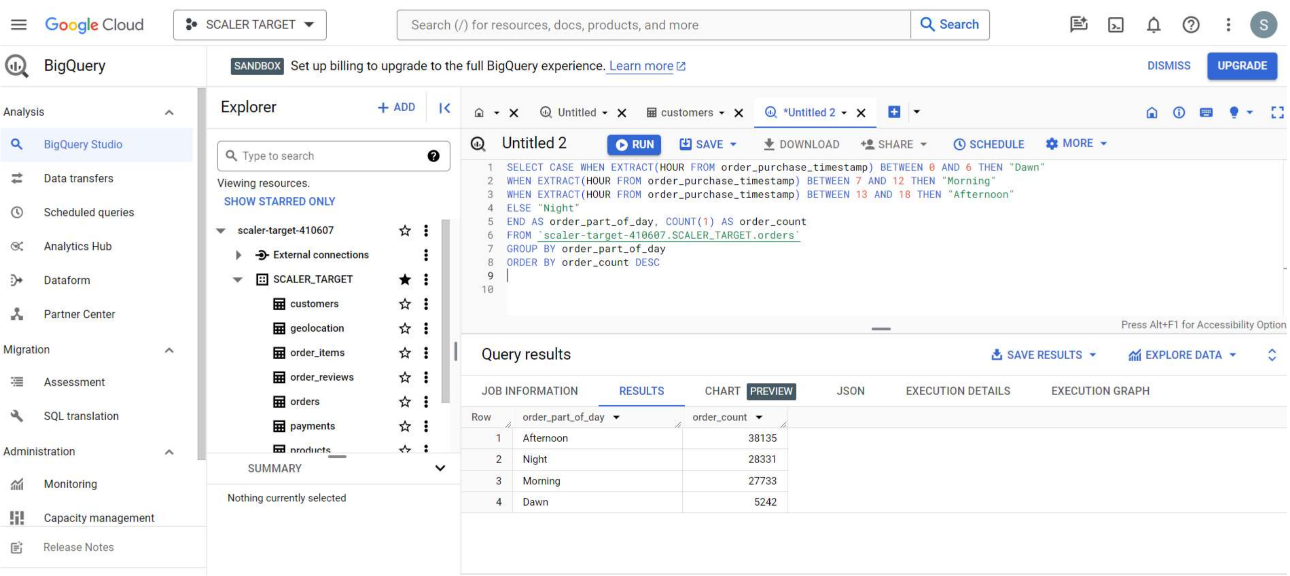
#Q3 - During what time of the day, do the Brazilian customers mostly place their orders? (Dawn, Morning, Afternoon or Night)

0-6 hrs : Dawn
7-12 hrs : Mornings
13-18 hrs : Afternoon
19-23 hrs : Night

Query:

```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 order_part_of_day, COUNT(1) AS order_count
FROM `scaler-target-410607.SCALER_TARGET.orders`
GROUP BY order_part_of_day
ORDER BY order_count DESC
```

Output:



Insight & Recommendations

Based on the hour component of the timestamp, the query divides the order timestamps into distinct time groups (Dawn, Morning, Afternoon, Night). Next, the results are arranged according to the number of orders that fell within each time period.

Analyzing data allows us to pinpoint the preferred time for Brazilian clients to place orders, offering insights into their preferences. Notably, **afternoons emerge as a prime period** for online shopping, with customers frequently making larger purchases. Leveraging this information, we can optimize operations by strategically scheduling customer assistance and launching targeted marketing efforts during peak ordering times. Additionally, the data indicates **minimal purchases during dawn**. This knowledge enables more efficient and tailored business strategies, aligning services with observed customer behaviour patterns.

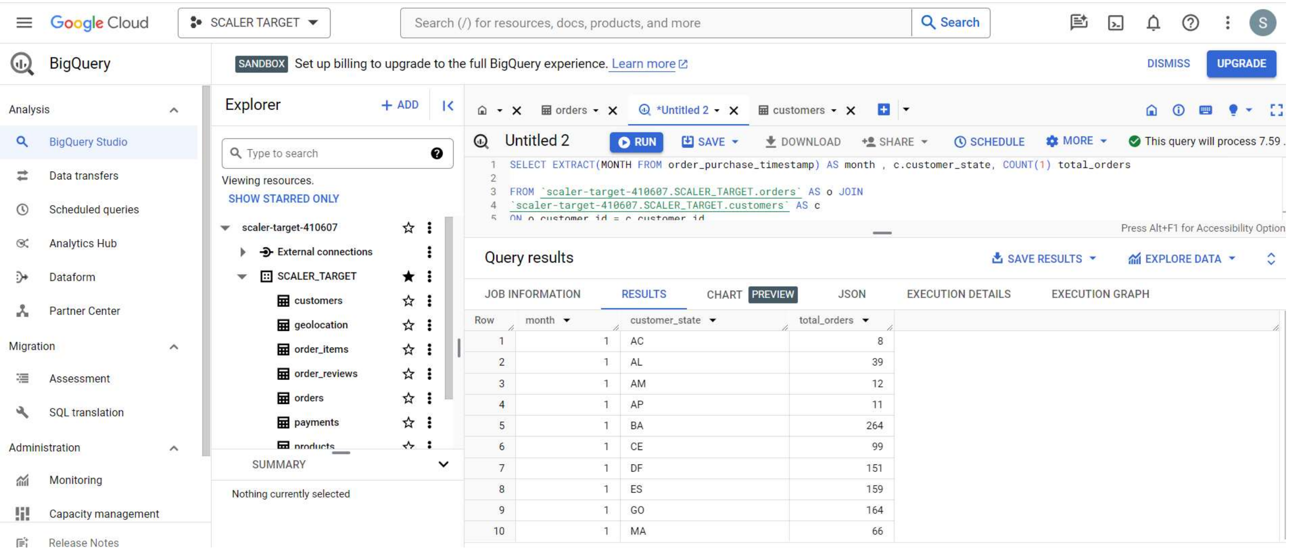
Task 3: Evolution of E-commerce orders in the Brazil region:

Q1 - Get the month on month no. of orders placed in each state.

Query:

```sql
SELECT EXTRACT(MONTH FROM o.order_purchase_timestamp) AS
month , c.customer_state, COUNT(1) total_orders

FROM `scaler-target-410607.SCALER_TARGET.orders` AS o JOIN
`scaler-target-410607.SCALER_TARGET.customers` AS c
ON o.customer_id = c.customer_id
GROUP BY month,c.customer_state
ORDER BY month,c.customer_state
```

Output:



Insight & Recommendations

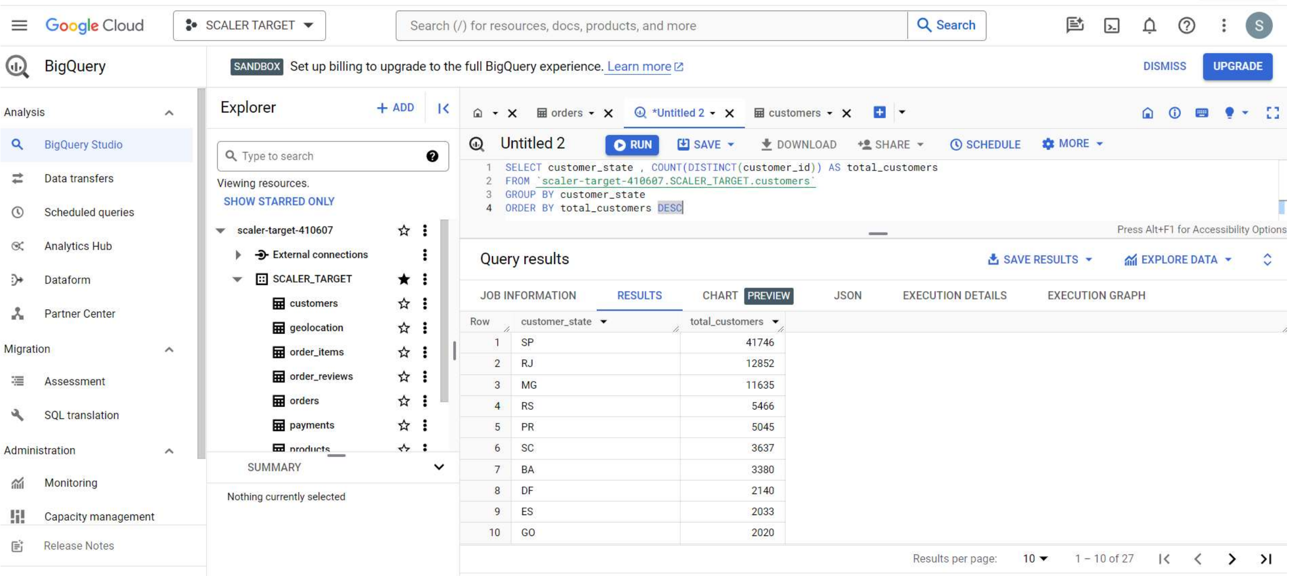
Examining query results reveals valuable insights into monthly order counts for each state. Detecting trends, patterns, and seasonality in order volumes over time allows us to identify states with consistently high orders. Notably, in our data, the state **SP** consistently records the **highest monthly orders**. Leveraging this information, we can tailor marketing efforts to states experiencing rising order volumes, address operational challenges in those with declining orders, and optimize inventory management based on state-specific order trends. These data-driven strategies enhance overall business efficiency and responsiveness to varying regional demands.

#Q2 - How are the customers distributed across all the states?

Query:

```
SELECT customer_state , COUNT(DISTINCT(customer_id)) AS
total_customers

FROM `scaler-target-410607.SCALER_TARGET.customers`
GROUP BY customer_state
ORDER BY total_customers DESC
```

Output:



Insight & Recommendations

Analyzing query results reveals client distribution across states, highlighting states with the highest and lowest customer numbers. Notably, **State SP has the most clients**, while **State RR has the fewest**. This information is crucial for market targeting, identifying expansion opportunities, and optimizing customer service strategies. Understanding client distribution aids in making informed decisions to enhance market reach and improve customer engagement.

Another point of analyzing customer distribution between states informs strategic decisions, helping identify growth areas and optimize company strategy for optimal results.

Task 4: Impact on Economy: Analyze the money movement by e-commerce by looking at order prices, freight and others.

Q1 - Get the % increase in the cost of orders from year 2017 to 2018 (include months between Jan to Aug only).

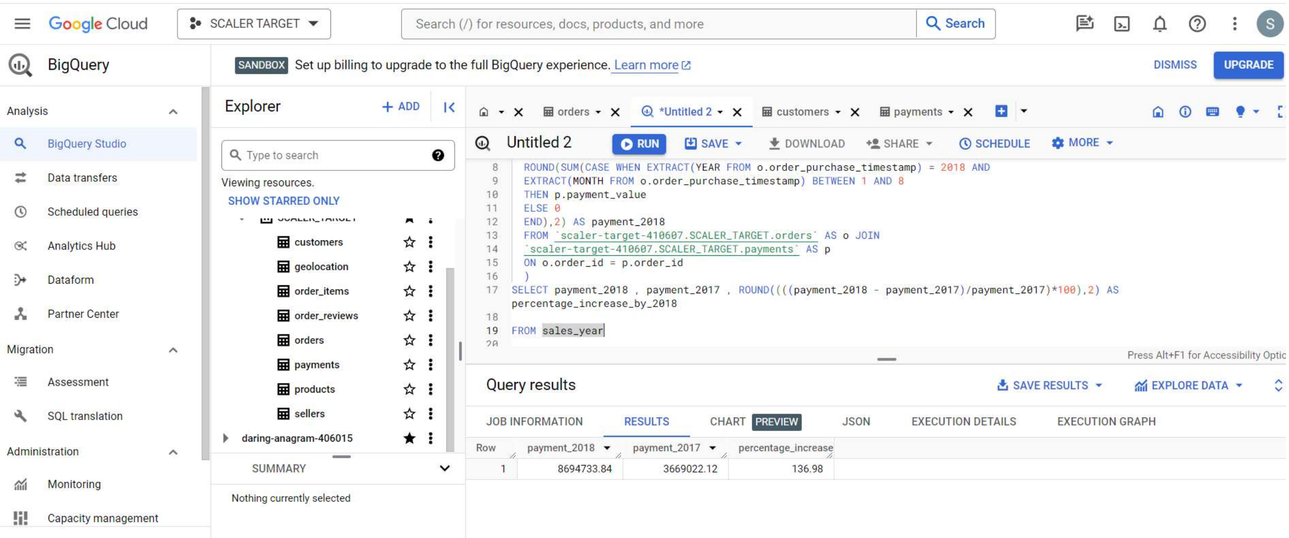
You can use the "payment_value" column in the payments table to get the cost of orders.

Query:

```
WITH sales_year AS(

 SELECT
 ROUND(SUM(CASE WHEN EXTRACT(YEAR FROM
o.order_purchase_timestamp) = 2017 AND
 EXTRACT(MONTH FROM o.order_purchase_timestamp) BETWEEN 1
AND 8
 THEN p.payment_value
 ELSE 0
```

```
 END),2) AS payment_2017 ,
 ROUND(SUM(CASE WHEN EXTRACT(YEAR FROM
o.order_purchase_timestamp) = 2018 AND
 EXTRACT(MONTH FROM o.order_purchase_timestamp) BETWEEN 1
AND 8
 THEN p.payment_value
 ELSE 0
 END),2) AS payment_2018
 FROM `scaler-target-410607.SCALER_TARGET.orders` AS o
JOIN
 `scaler-target-410607.SCALER_TARGET.payments` AS p
 ON o.order_id = p.order_id
 )
SELECT payment_2018 , payment_2017 , ROUND((((payment_2018
- payment_2017)/payment_2017)*100),2) AS
percentage_increase_by_2018

FROM sales_year
```

Output:



Insight & Recommendations

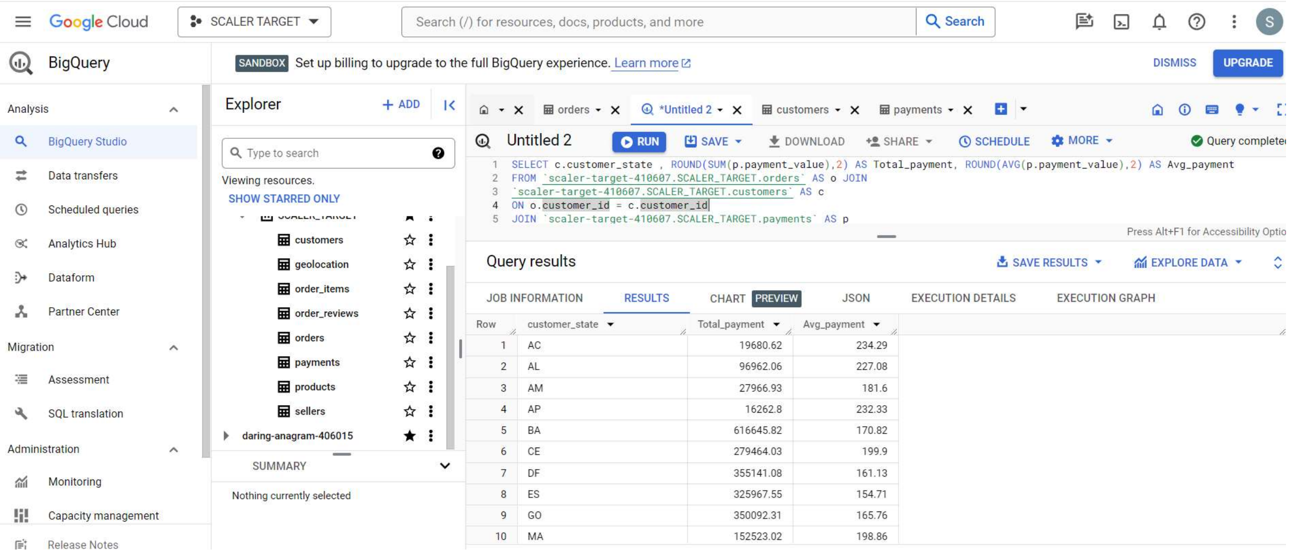
Only orders placed from **January to August** are taken into consideration for the years **2017** and **2018**. The overall order amount for the year **2017 is 3669022.12**, and for the **year 2018**, **it is 8694733.84**. The query examines the **monthly prices** between 2017 and 2018 to determine the percentage increase. According to the data, there was a **growth rate** of **nearly 137% from 2017 to 2018**.

Q2 - Calculate the Total & Average value of order price for each state.

Query:

```sql
SELECT c.customer_state , ROUND(SUM(p.payment_value),2) AS
Total_payment, ROUND(AVG(p.payment_value),2) AS
Avg_payment

FROM `scaler-target-410607.SCALER_TARGET.orders` AS o JOIN
`scaler-target-410607.SCALER_TARGET.customers` AS c
ON o.customer_id = c.customer_id
JOIN `scaler-target-410607.SCALER_TARGET.payments` AS p
ON o.order_id = p.order_id
GROUP BY c.customer_state
ORDER BY c.customer_state ASC
```

Output:



Insight & Recommendations

The t**otal order prices** for each state are shown in the "**total_order_price**" column, which also shows the total number of orders placed. In addition, details regarding the **average order value** for every state can be found in the "**average_order_price**" column.

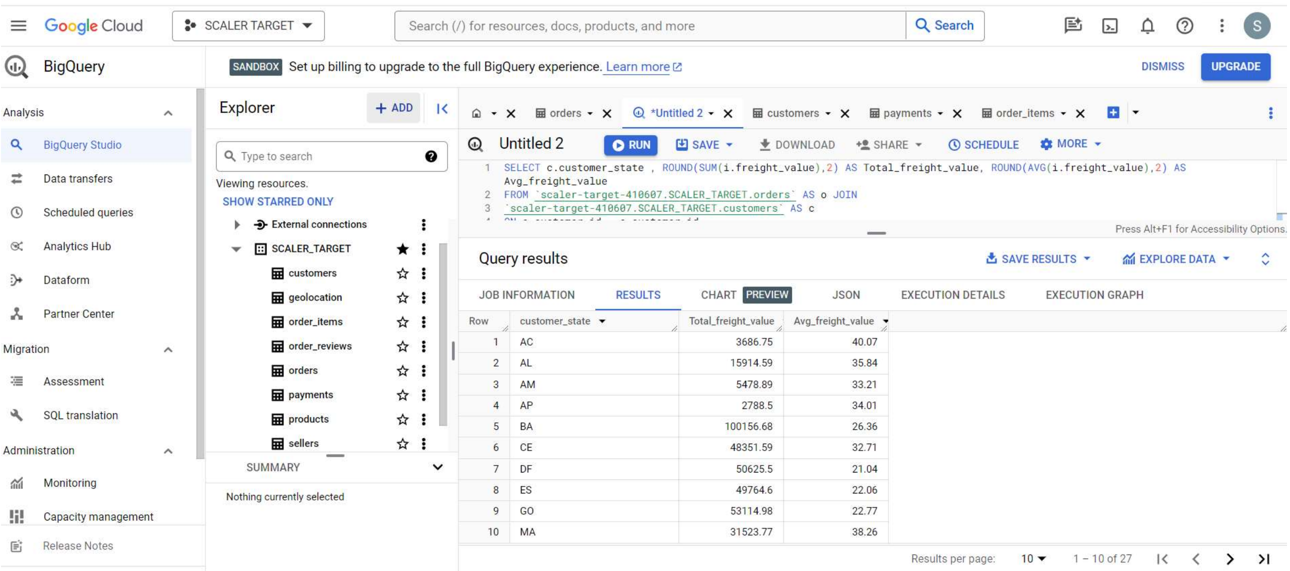
By examining these findings, states with sizeable **total order** values can be found, suggesting **potentially profitable marketplaces**. Comparing **average order** costs across states can assist in identifying **locations with different purchasing patterns**, which can then be used to design **customized marketing or pricing strategies**. However, it is essential to consider each state's specifics to have a thorough understanding and make informed choices.

Q3 - Calculate the Total & Average value of order freight for each state.

Query:

```sql
SELECT c.customer_state , ROUND(SUM(i.freight_value),2) AS
Total_freight_value, ROUND(AVG(i.freight_value),2) AS
Avg_freight_value

FROM `scaler-target-410607.SCALER_TARGET.orders` AS o JOIN
`scaler-target-410607.SCALER_TARGET.customers` AS c
ON o.customer_id = c.customer_id
JOIN `scaler-target-410607.SCALER_TARGET.order_items` AS i
ON o.order_id = i.order_id
GROUP BY c.customer_state
ORDER BY c.customer_state ASC
```

Output:



Insight & Recommendations

Data analysis reveals states like **SP with highest total freight costs** and **RR with the lowest costs**, yet **RR's average freight cost** is **nearly three times higher than SP**. This indicates regions potentially facing l**ogistical challenges** or higher shipping prices. Analyzing these findings is crucial for optimizing l**ogistics operations** and **pricing strategies**. A comparative study of average order freight costs across states enables the identification of areas with varying shipping expenses, aiding strategic decision-making. Understanding differences in order freight rates provides insights into **local shipping practices**, supplier locations, and **customer preferences**. This knowledge proves invaluable for optimizing processes and **cutting costs**, thereby **enhancing** overall **business efficiency**.

Task 5: **Analysis based on sales, freight and delivery time.**

Q1 - 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.

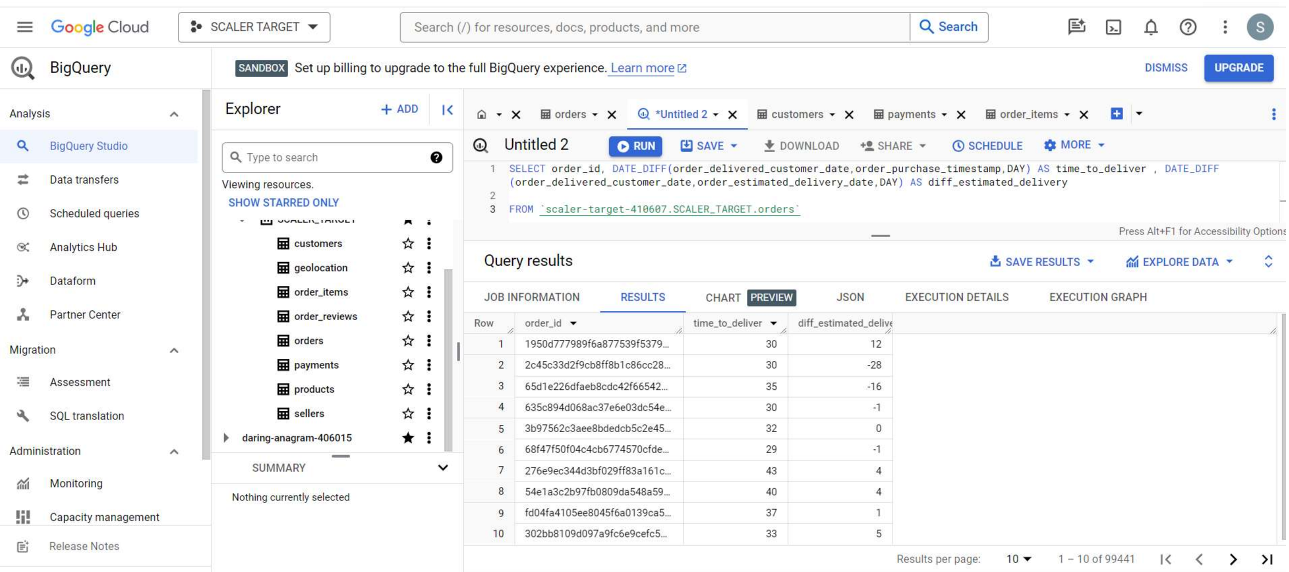
You can calculate the delivery time and the difference between the estimated & actual delivery date using the given formula:

- **time_to_deliver** = order_delivered_customer_date - order_purchase_timestamp
- **diff_estimated_delivery** = order_delivered_customer_date – order_estimated_delivery_date

Query:

```
SELECT order_id,
DATE_DIFF(order_delivered_customer_date,order_purchase_timestamp,DAY) AS time_to_deliver ,
DATE_DIFF(order_delivered_customer_date,order_estimated_delivery_date,DAY) AS diff_estimated_delivery

FROM `scaler-target-410607.SCALER_TARGET.orders`
```

Output:



Insight & Recommendations

Delivery time and diff_estimated_delivery analysis sheds light on any **delays** or **early deliveries** relative to the scheduled timetable, offering insightful information about how well the delivery process is working.

Finding **patterns** and **anomalies** in these columns might be useful in **determining** what **influences delivery timelines** or causes **differences** between **projected** and **actual dates**. By utilising these insights, one may **optimise delivery processes**, manage customer expectations more effectively, increase customer happiness, and **improve logistics operations efficiency.**

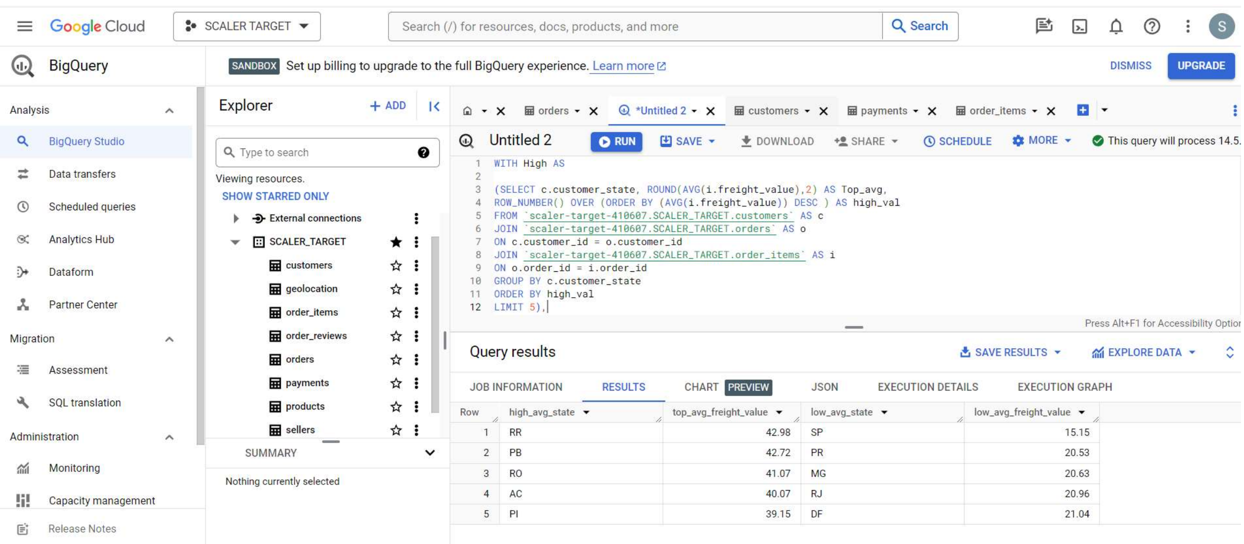
Q2 - Find out the top 5 states with the highest & lowest average freight value.

Query:

```
WITH High AS

(SELECT c.customer_state, ROUND(AVG(i.freight_value),2) AS
Top_avg,
ROW_NUMBER() OVER (ORDER BY (AVG(i.freight_value)) DESC )
AS high_val
FROM `scaler-target-410607.SCALER_TARGET.customers` AS c
JOIN `scaler-target-410607.SCALER_TARGET.orders` AS o
ON c.customer_id = o.customer_id
JOIN `scaler-target-410607.SCALER_TARGET.order_items` AS i
ON o.order_id = i.order_id
GROUP BY c.customer_state
```

```
ORDER BY high_val
LIMIT 5),
Low AS
(SELECT c.customer_state, ROUND(AVG(i.freight_value),2) AS
Bot_avg,
ROW_NUMBER() OVER (ORDER BY (AVG(i.freight_value)) ) AS
bot_val
FROM `scaler-target-410607.SCALER_TARGET.customers` AS c
JOIN `scaler-target-410607.SCALER_TARGET.orders` AS o
ON c.customer_id = o.customer_id
JOIN `scaler-target-410607.SCALER_TARGET.order_items` AS i
ON o.order_id = i.order_id
GROUP BY c.customer_state
ORDER BY bot_val
LIMIT 5
)

SELECT
h.customer_state AS high_avg_state, h.Top_avg AS
top_avg_freight_value,
l.customer_state AS low_avg_state, l.Bot_avg AS
low_avg_freight_value
FROM High AS h
JOIN Low AS l
ON h.high_val = l.bot_val
```

Output:



Insight & Recommendations

States like **RR and PB** that have **high average freight values** may see higher shipping prices as a result of supply chain complexity, remote locations, or higher transportation costs. We may save expenses and improve logistical operations for our business by locating regions with **lower average freight values**, such as states like **PR and SP**. This information helps us identify possibilities to reduce costs in our supply chain operations, negotiate freight charges, and develop targeted efforts.

Nevertheless, it's important to take into account other factors, like carrier availability, distance, transportation infrastructure, and regional economic variances, when interpreting these findings. These elements are essential to comprehending the **complexities of shipping dynamics** and guaranteeing a thorough strategy for reducing expenses and **logistics optimisation**. By taking these factors into consideration, we can improve our **negotiating position**, hone our strategy, and make well-informed decisions that will increase the overall **effectiveness** of our supply chain operations.

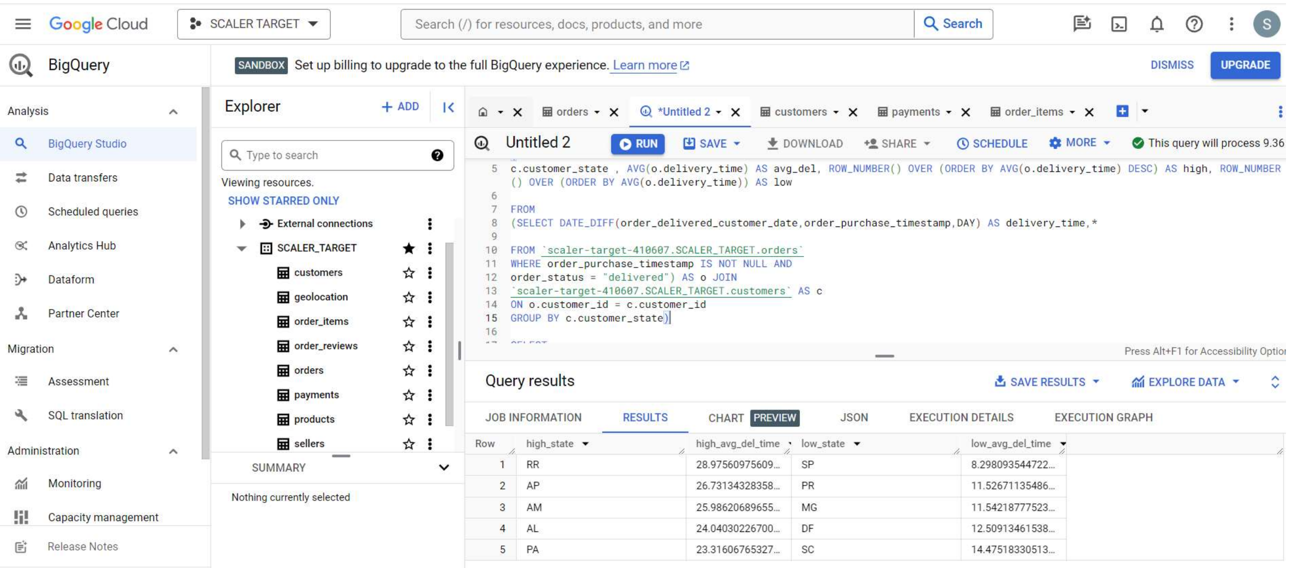
Q3 - Find out the top 5 states with the highest & lowest average delivery time.

Query:

```
WITH avg_delivery AS

(SELECT
c.customer_state , AVG(o.delivery_time) AS avg_del,
ROW_NUMBER() OVER (ORDER BY AVG(o.delivery_time) DESC) AS
high, ROW_NUMBER() OVER (ORDER BY AVG(o.delivery_time)) AS
low
```

```
FROM
(SELECT
DATE_DIFF(order_delivered_customer_date,order_purchase_tim
estamp,DAY) AS delivery_time,*

FROM `scaler-target-410607.SCALER_TARGET.orders`
WHERE order_purchase_timestamp IS NOT NULL AND
order_status = "delivered") AS o JOIN
`scaler-target-410607.SCALER_TARGET.customers` AS c
ON o.customer_id = c.customer_id
GROUP BY c.customer_state)

SELECT
a1.customer_state AS high_state , ROUND(a1.avg_del,2) AS
high_avg_del_time ,
a2.customer_state AS low_state ,ROUND(a2.avg_del,2) AS
low_avg_del_time
FROM avg_delivery AS a1 JOIN
avg_delivery AS a2
ON a1.high = a2.low
LIMIT 5
```

<u>Output:</u>



Insight & Recommendations

States like **SP and PR with the lowest** average delivery times are compared against states like **RR and AP with the highest** average delivery times in order to determine whether regions have effective delivery operations, shorter transit times, and **strong logistics networks**. For our business, which seeks to i**mprove customer satisfaction**, operational effectiveness, and **delivery process optimisation**, these insights are priceless. It becomes easier to set reasonable expectations for customers based on local delivery time norms.

It is crucial to take into account additional elements like **population density**, the distinction between **urban** and **rural** areas, consumer expectations, and certain logistical constraints when analysing the data and drawing inferences from these insights. By **tailoring strategies based** on **regional variations** and **specific challenges**, our company can better align its delivery services with customer expectations, ultimately fostering improved satisfaction and operational effectiveness.

Q4 - Find out the top 5 states where the order delivery is really fast as compared to the estimated date of delivery.

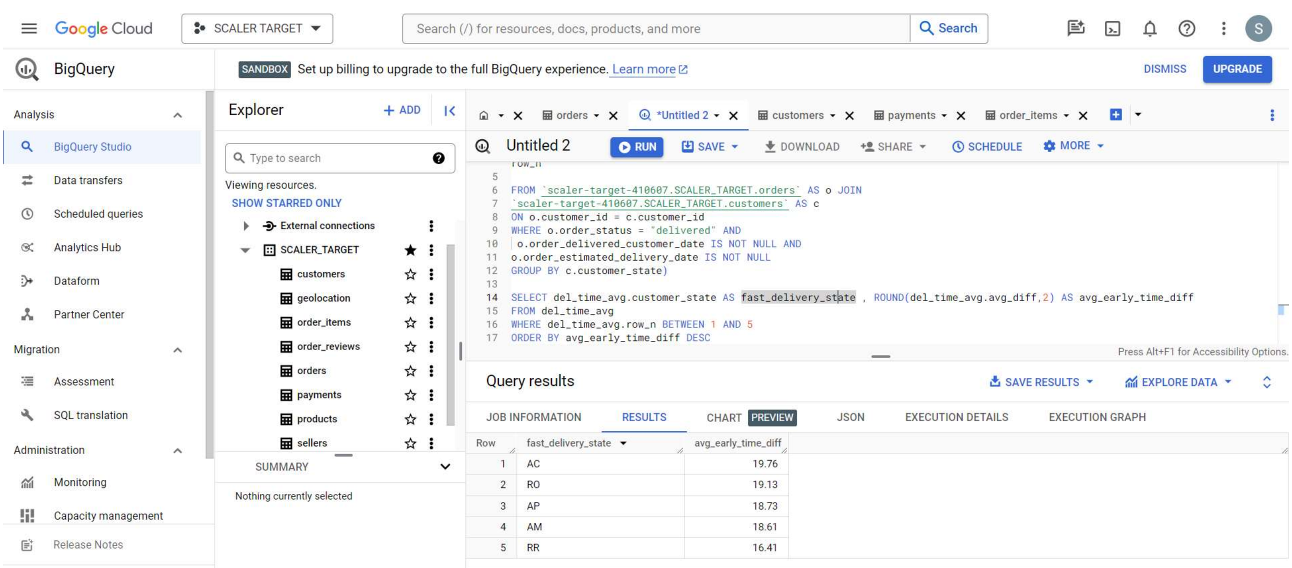
You can use the difference between the averages of actual & estimated delivery date to figure out how fast the delivery was for each state.

Query:

```
WITH del_time_avg AS

(SELECT c.customer_state ,
```

```
AVG(DATE_DIFF(o.order_estimated_delivery_date,o.order_delivered_customer_date,day)) AS avg_diff, ROW_NUMBER() OVER
(ORDER BY
AVG(DATE_DIFF(o.order_estimated_delivery_date,o.order_delivered_customer_date,day)) DESC ) AS row_n

FROM `scaler-target-410607.SCALER_TARGET.orders` AS o JOIN
`scaler-target-410607.SCALER_TARGET.customers` AS c
ON o.customer_id = c.customer_id
WHERE o.order_status = "delivered" AND
 o.order_delivered_customer_date IS NOT NULL AND
o.order_estimated_delivery_date IS NOT NULL
GROUP BY c.customer_state)

SELECT del_time_avg.customer_state AS
fast_delivery_state , ROUND(del_time_avg.avg_diff,2) AS
avg_early_time_diff
FROM del_time_avg
WHERE del_time_avg.row_n BETWEEN 1 AND 5
ORDER BY avg_early_time_diff DESC
```

Output:



Insight & Recommendations

Our organisation can take advantage of quicker delivery times to establish a reputation for prompt and dependable service because we operate in states with the h**ighest average delivery speeds**, such as **AC,RO,AP,AM and RR.** Showcasing this effectiveness can draw in more business and greatly raise client happiness. These insights are useful instruments for streamlining operations in general, streamlining logistics, and pinpointing prospective growth areas where expedited order fulfilment has worked well.

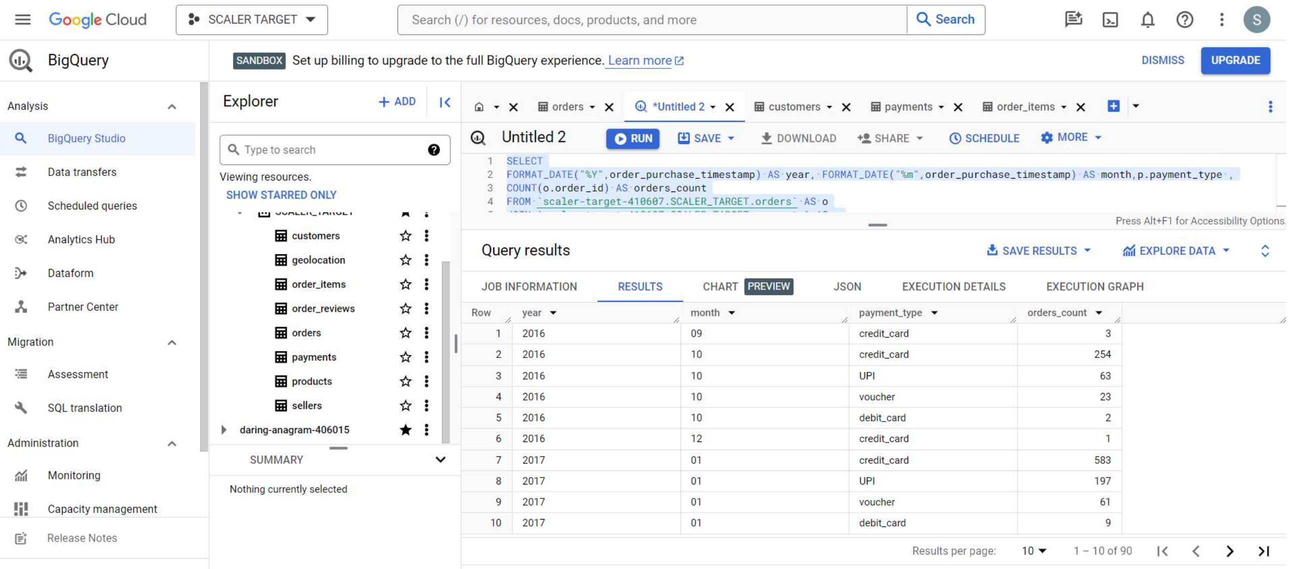
Task 6: Analysis based on the payments:

Q1 - Find the month on month no. of orders placed using different payment types.

Query:

```
SELECT FORMAT_DATE("%Y",order_purchase_timestamp) AS year,
FORMAT_DATE("%m",order_purchase_timestamp) AS
month,p.payment_type ,

COUNT(o.order_id) AS orders_count
FROM `scaler-target-410607.SCALER_TARGET.orders` AS o
JOIN `scaler-target-410607.SCALER_TARGET.payments` AS p
ON o.order_id = p.order_id
GROUP BY p.payment_type,year,month
ORDER BY year,month
```

Output:



Insight & Recommendations

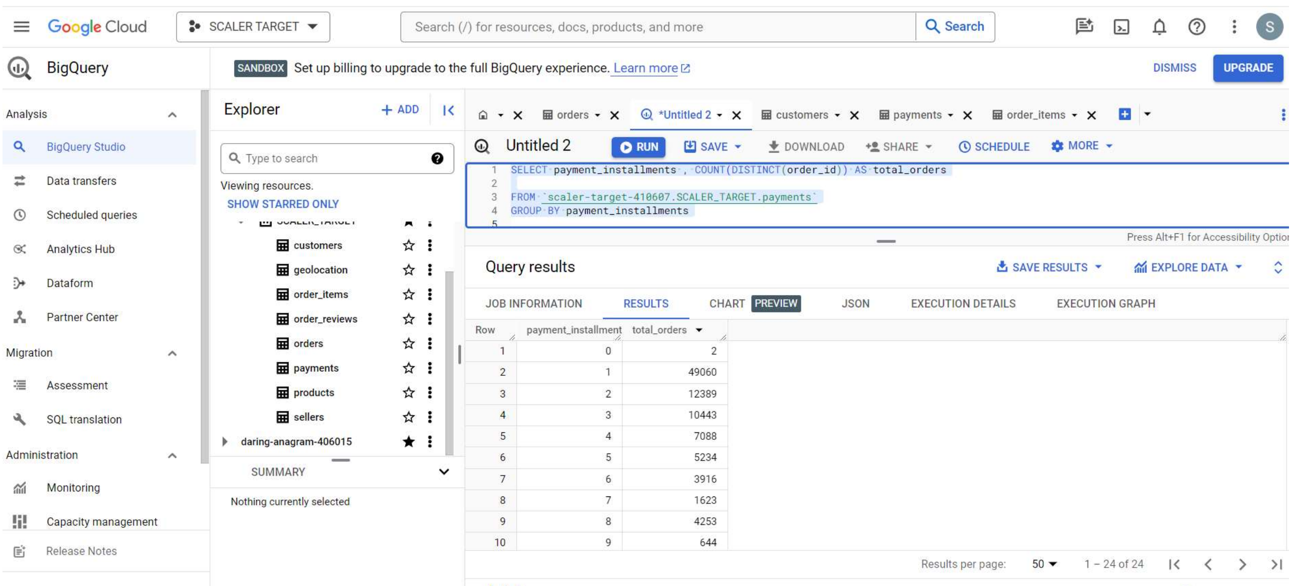
We've found that **credit cards** are **most frequently** used for **order purchases. The** importance of credit card usage in November 2017 emphasises how important it is to monitor order count fluctuations from month to month. Understanding seasonality, determining peak months, and evaluating the influence of marketing initiatives or outside variables on consumer behaviour are all made easier with the help of this analysis. Businesses can optimise payment processes, customise marketing efforts, and improve overall customer experiences by utilising insights regarding payment preferences over multiple months. This helps firms align their business strategy, marketing offers (e.g. offers on credit card) with the behaviours of their observed consumer base.

Q2 - Find the no. of orders placed on the basis of the payment installments that have been paid.

Query:

```
SELECT payment_installments , COUNT(DISTINCT(order_id)) AS total_orders

FROM `scaler-target-410607.SCALER_TARGET.payments`
GROUP BY payment_installments
```

Output:



Insight & Recommendations

A total of **49,060 orders** were made with a **single payment installment.** This analysis serves to assess the popularity and preference for payment installment alternatives among clients. Observing whether customers tend to choose a specific number of payment installments can provide insights into their preferences for budgeting or financing options. By monitoring the distribution of orders based on payment installments, the most flexible and preferred payment methods can be revealed.

- SABYASACHI BANERJEE

Batch:DSML Nov23 Beginner Tue

NOTE: "Google BigQuery" is used to solve all of the questions.