**PROJECT 5 - TARGET**

**Context**

Target is one of the world’s most recognized brands and one of America’s leading retailers. Target makes itself a preferred shopping destination by offering outstanding value, inspiration, innovation and an exceptional guest experience that no other retailer can deliver.

This business case has information of 100k orders from 2016 to 2018 made at Target in Brazil. Its features allow viewing an order from multiple dimensions: from order status, price, payment and freight performance to customer location, product attributes and finally reviews written by customers.

The **customers.csv** contain following features:

|  |  |
| --- | --- |
| **Features** | **Description** |
| customer\_id | Id of the consumer who made the purchase. |
| customer\_unique\_id | Unique Id of the consumer. |
| customer\_zip\_code\_prefix | Zip Code of the location of the consumer. |
| customer\_city | Name of the City from where order is made. |
| customer\_state | State Code from where order is made(Ex- sao paulo-SP). |

The **sellers.csv** contains following features:

|  |  |
| --- | --- |
| **Features** | **Description** |
| seller\_id | Unique Id of the seller registered |
| seller\_zip\_code\_prefix | Zip Code of the location of the seller. |
| seller\_city | Name of the City of the seller. |
| seller\_state | State Code (Ex- sao paulo-SP) |

The **order\_items.csv** contain following features:

|  |  |
| --- | --- |
| **Features** | **Description** |
| order\_id | A unique id of order made by the consumers. |
| order\_item\_id | A Unique id given to each item ordered in the order. |
| product\_id | A unique id given to each product available on the site. |
| seller\_id | Unique Id of the seller registered in Target. |
| shipping\_limit\_date | The date before which shipping of the ordered product must be completed. |
| price | Actual price of the products ordered . |
| freight\_value | Price rate at which a product is delivered from one point to another. |

The **payments.csv** contain following features:

|  |  |
| --- | --- |
| **Features** | **Description** |
| order\_id | A unique id of order made by the consumers. |
| payment\_sequential | sequences of the payments made in case of EMI. |
| payment\_type | mode of payment used.(Ex-Credit Card) |
| payment\_installments | number of installments in case of EMI purchase. |
| payment\_value | Total amount paid for the purchase order. |

The **orders.csv** contain following features:

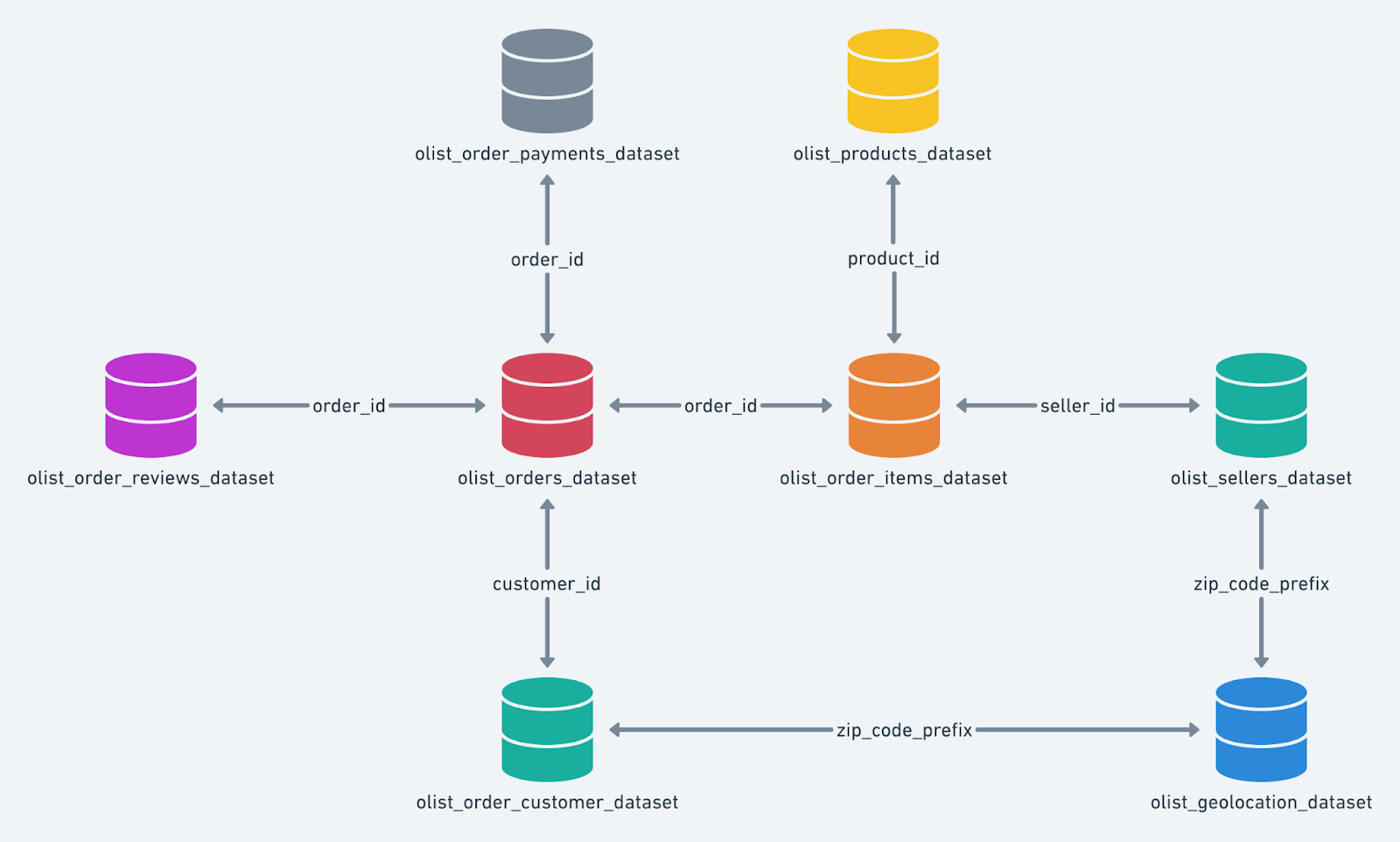
|  |  |
| --- | --- |
| **Features** | **Description** |
| order\_id | A unique id of order made by the consumers. |
| customer\_id | Id of the consumer who made the purchase. |
| order\_status | status of the order made i.e delivered, shipped etc. |
| order\_purchase\_timestamp | Timestamp of the purchase. |
| order\_delivered\_carrier\_date | delivery date at which carrier made the delivery. |
| order\_delivered\_customer\_date | date at which customer got the product. |
| order\_estimated\_delivery\_date | estimated delivery date of the products. |

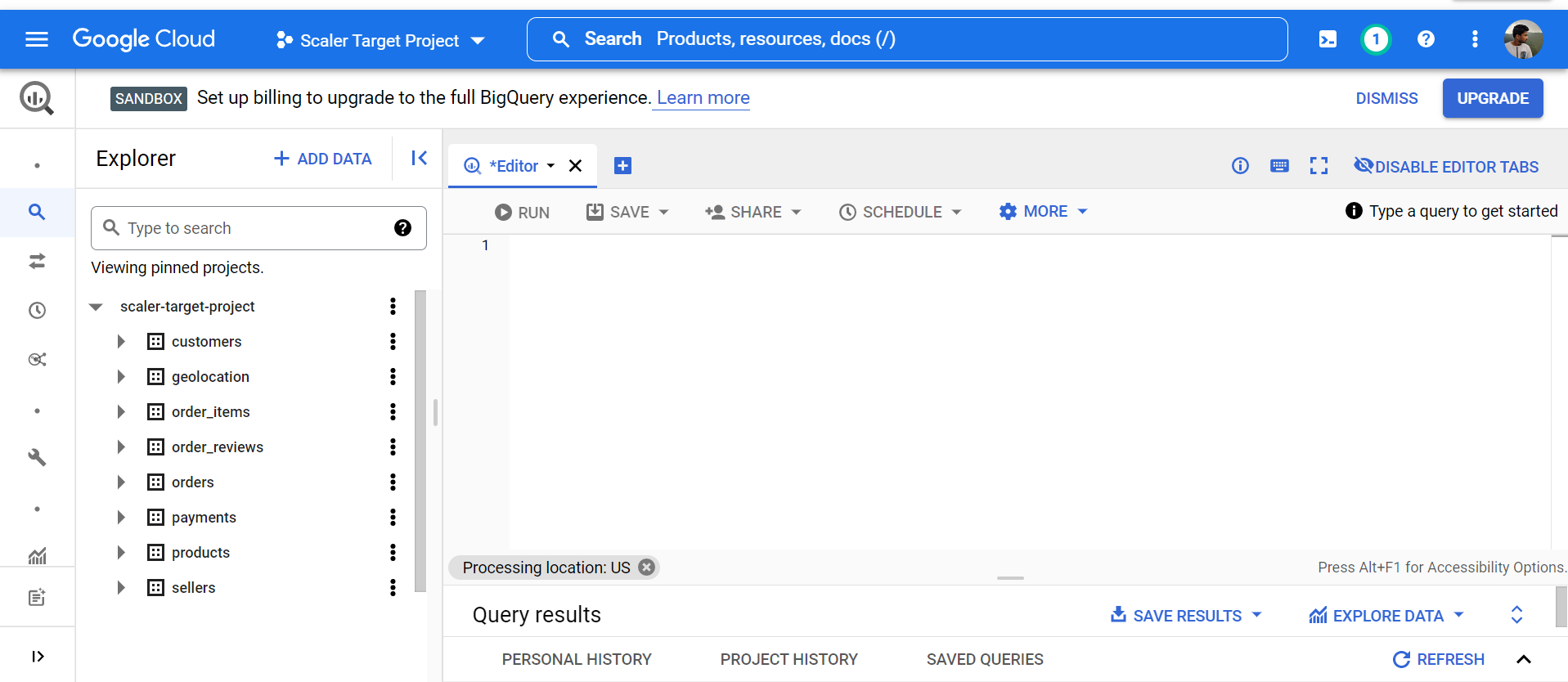
The **reviews.csv** contain following features:

|  |  |
| --- | --- |
| **Features** | **Description** |
| review\_id | Id of the review given on the product ordered by the order id. |
| order\_id | A unique id of order made by the consumers. |
| review\_score | review score given by the customer for each order on the scale of 1–5. |
| review\_comment\_title | Title of the review |
| review\_comment\_message | Review comments posted by the consumer for each order. |
| review\_creation\_date | Timestamp of the review when it is created. |
| review\_answer\_timestamp | Timestamp of the review answered. |

The **products.csv** contain following features:

|  |  |
| --- | --- |
| **Features** | **Description** |
| product\_id | A unique identifier for the proposed project. |
| product\_category\_name | Name of the product category |
| product\_name\_lenght | length of the string which specifies the name given to the products ordered. |
| product\_description\_lenght | length of the description written for each product ordered on the site. |
| product\_photos\_qty | Number of photos of each product ordered available on the shopping portal. |
| product\_weight\_g | Weight of the products ordered in grams. |
| product\_length\_cm | Length of the products ordered in centimeters. |
| product\_height\_cm | Height of the products ordered in centimeters. |
| product\_width\_cm | width of the product ordered in centimeters. |

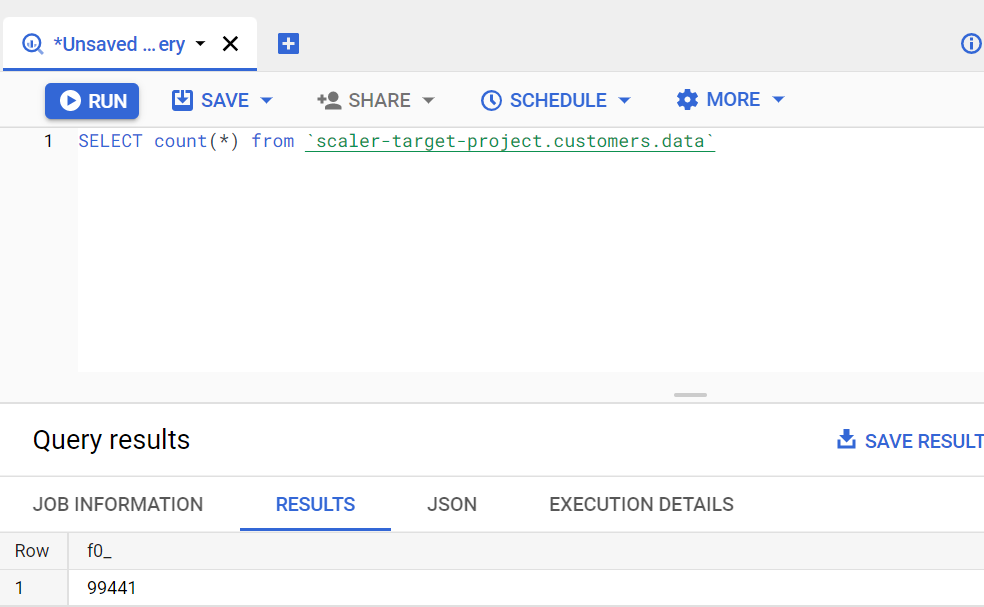




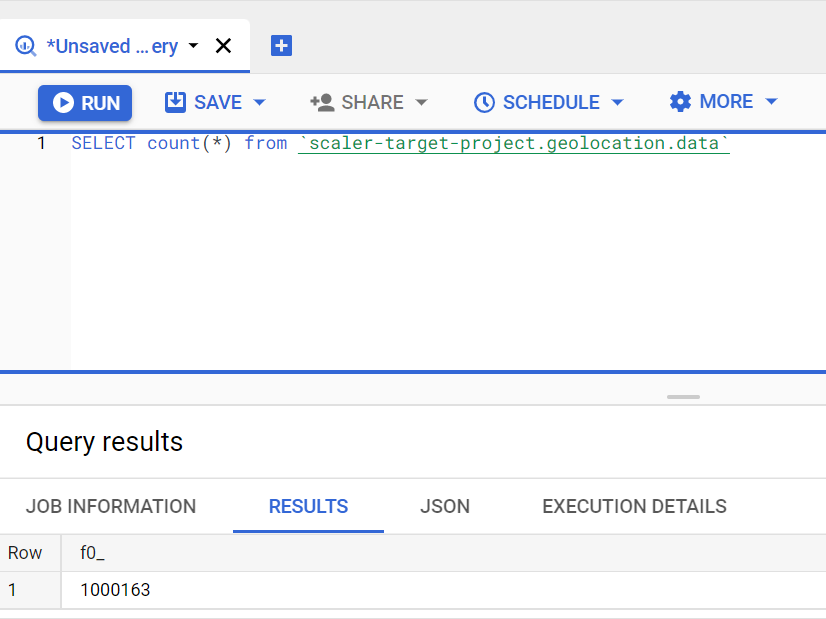
QUESTIONS:

1. Import the dataset and do usual exploratory analysis steps like checking the structure & characteristics of the dataset  
   1. Get number of rows in the data

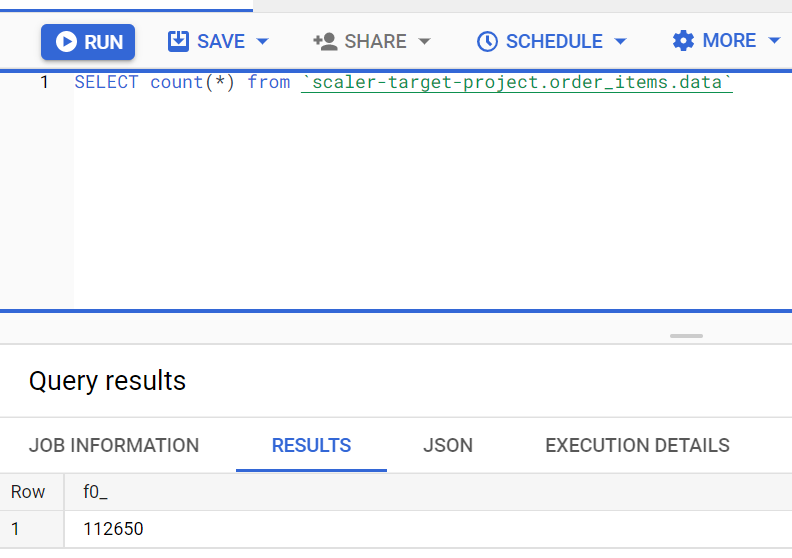
* **Customers Table** – 99441 Rows



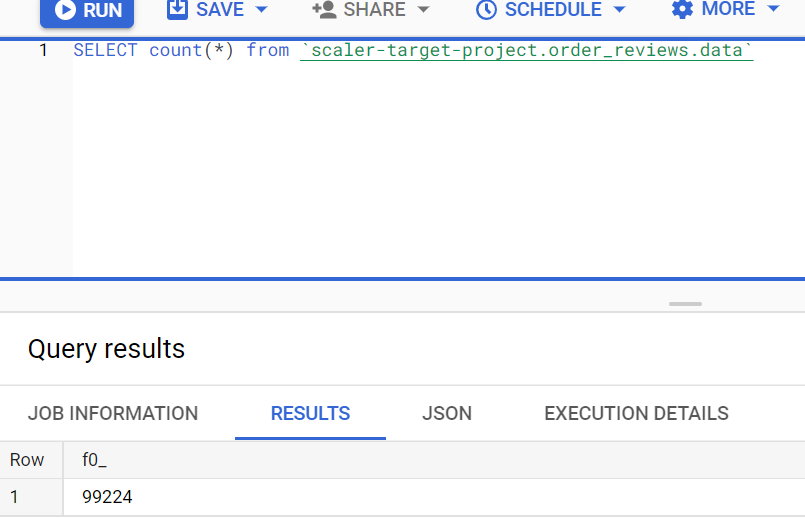
* **Geolocation Table** – 1000163 Rows



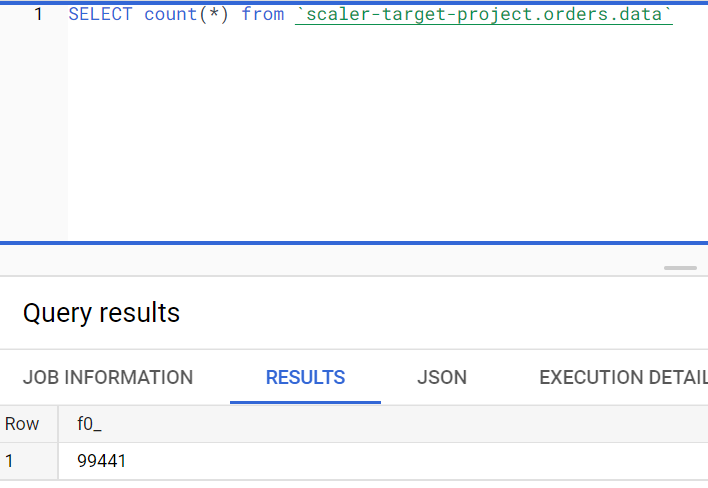
* **Order\_Items Table** –112650 Rows



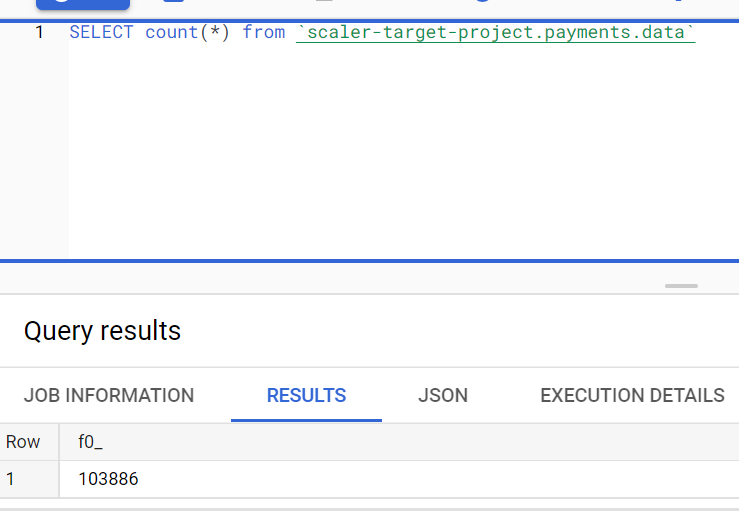
* **Order\_Reviews Table** – 99224 Rows



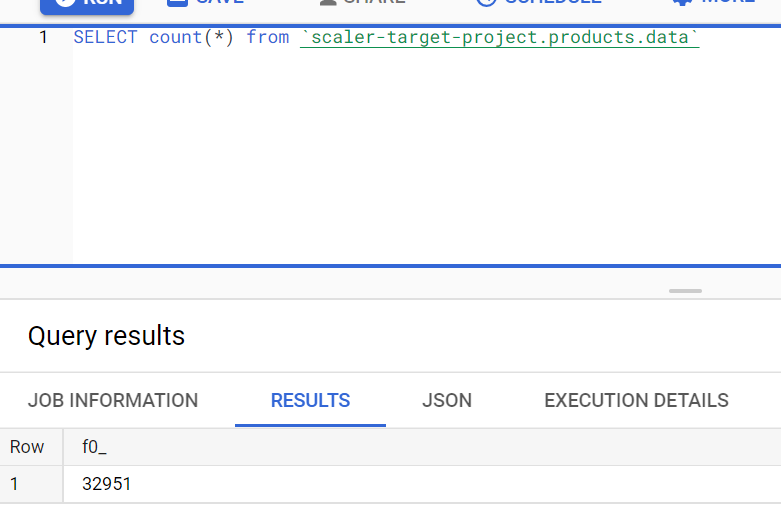
* **Orders Table** – 99441 Rows



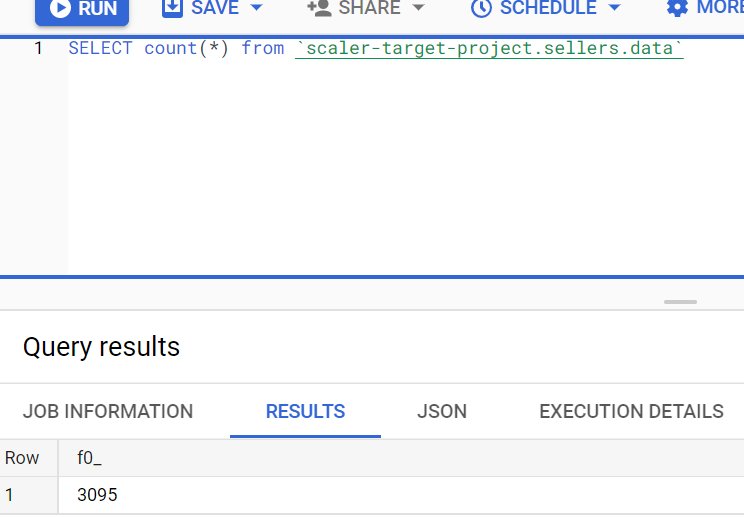
* **Payments Table** – 103886 Rows



* **Products Table** – 32951 Rows



* **Sellers Table** – 3095 Rows



* 1. Number of null or missing values in a column

For Customers Table :

1. SELECT count(\*) from `scaler-target-project.customers.data` where customer\_id IS NULL; 0 Null Values
2. SELECT count(\*) from `scaler-target-project.customers.data` where customer\_unique\_id IS NULL; 0 Null Values
3. SELECT count(\*) from `scaler-target-project.customers.data` where customer\_zip\_code\_prefix IS NULL; 0 Null Values
4. SELECT count(\*) from `scaler-target-project.customers.data` where customer\_city IS NULL; 0 Null Values
5. SELECT count(\*) from `scaler-target-project.customers.data` where customer\_state IS NULL; 0 Null Values

For Sellers Table :

1. SELECT count(\*) from `scaler-target-project.sellers.data` where seller\_id IS NULL;

0 Null Values

1. SELECT count(\*) from `scaler-target-project.sellers.data` where seller\_zip\_code\_prefix IS NULL; 0 Null Values
2. SELECT count(\*) from `scaler-target-project.sellers.data` where seller\_city IS NULL;

0 Null Values

1. SELECT count(\*) from `scaler-target-project.sellers.data` where seller\_state IS NULL;

0 Null Values

For Order\_Items Table :

1. SELECT count(\*) from `scaler-target-project.order\_items.data` where order\_id IS NULL

0 Null Values

1. SELECT count(\*) from `scaler-target-project.order\_items.data` where order\_item\_id IS NULL 0 Null Values
2. SELECT count(\*) from `scaler-target-project.order\_items.data` where product\_id IS NULL 0 Null Values
3. SELECT count(\*) from `scaler-target-project.order\_items.data` where seller\_id IS NULL

0 Null Values

1. SELECT count(\*) from `scaler-target-project.order\_items.data` where shipping\_limit\_date IS NULL 0 Null Values
2. SELECT count(\*) from `scaler-target-project.order\_items.data` where price IS NULL

0 Null Values

1. SELECT count(\*) from `scaler-target-project.order\_items.data` where freight\_value IS NULL 0 Null Values

For Payments Table :

1. SELECT count(\*) from `scaler-target-project.payments.data` where order\_id IS NULL;

0 Null Values

1. SELECT count(\*) from `scaler-target-project.payments.data` where payment\_sequential IS NULL; 0 Null Values
2. SELECT count(\*) from `scaler-target-project.payments.data` where payment\_type IS NULL; 0 Null Values
3. SELECT count(\*) from `scaler-target-project.payments.data` where payment\_installments IS NULL; 0 Null Values
4. SELECT count(\*) from `scaler-target-project.payments.data` where payment\_type IS NULL; 0 Null Values

For Orders Table :

1. SELECT count(\*) from `scaler-target-project.orders.data` where order\_id IS NULL

0 Null Values

1. SELECT count(\*) from `scaler-target-project.orders.data` where customer\_id IS NULL

0 Null Values

1. SELECT count(\*) from `scaler-target-project.orders.data` where order\_status IS NULL

0 Null Values

1. SELECT count(\*) from `scaler-target-project.orders.data` where order\_purchase\_timestamp IS NULL 0 Null Values
2. SELECT count(\*) from `scaler-target-project.orders.data` where order\_delivered\_carrier\_date IS NULL 1783 Null Values
3. SELECT count(\*) from `scaler-target-project.orders.data` where order\_delivered\_customer\_date IS NULL 2965 Null Values
4. SELECT count(\*) from `scaler-target-project.orders.data` where order\_estimated\_delivery\_date IS NULL 0 Null Values

For Order\_Reviews Table :

1. SELECT count(\*) from `scaler-target-project.order\_reviews.data` where review\_id IS NULL; 0 Null Values
2. SELECT count(\*) from `scaler-target-project.order\_reviews.data` where order\_id IS NULL; 0 Null Values
3. SELECT count(\*) from `scaler-target-project.order\_reviews.data` where review\_score IS NULL; 0 Null Values
4. SELECT count(\*) from `scaler-target-project.order\_reviews.data` where review\_comment\_title IS NULL; 87675 Null Values
5. SELECT count(\*) from `scaler-target-project.order\_reviews.data` where review\_creation\_date IS NULL; 0 Null Values
6. SELECT count(\*) from `scaler-target-project.order\_reviews.data` where review\_answer\_timestamp IS NULL; 0 Null Values

For Products Table :

1. SELECT count(\*) from `scaler-target-project.products.data` where product\_id IS NULL

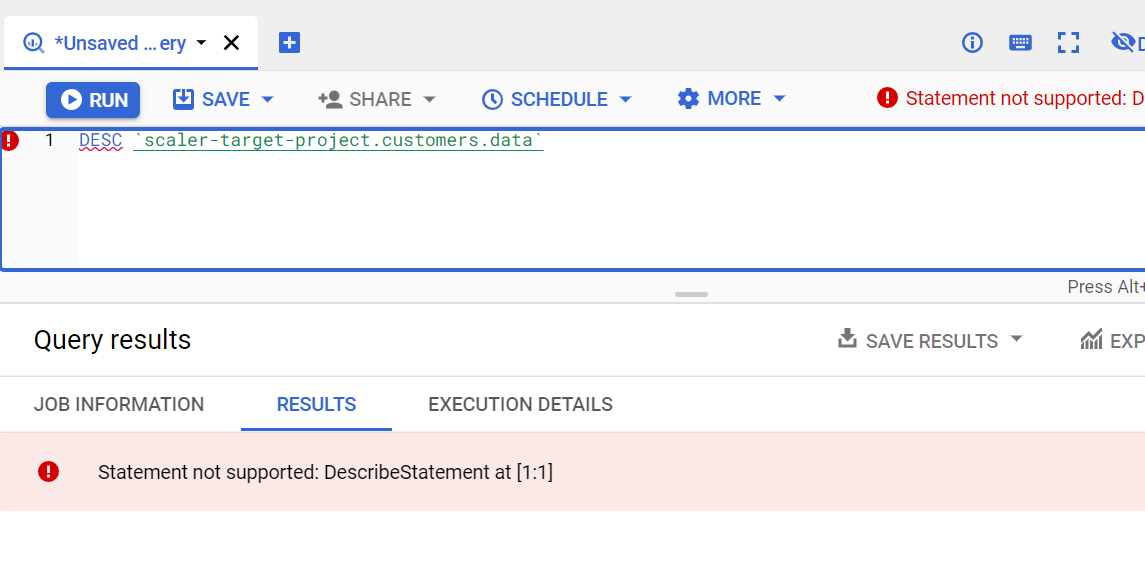
0 Null Values

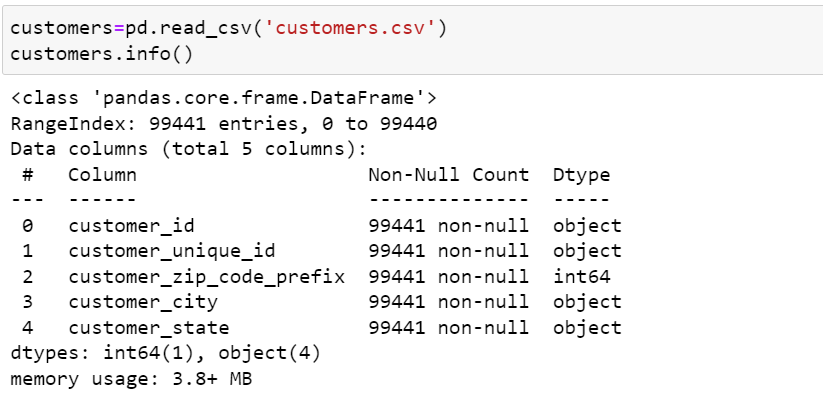
1. SELECT count(\*) from `scaler-target-project.products.data` where product\_category IS NULL 610 0 Null Values
2. SELECT count(\*) from `scaler-target-project.products.data` where product\_name\_length IS NULL 610 0 Null Values
3. SELECT count(\*) from `scaler-target-project.products.data` where product\_description\_length IS NULL 610 0 Null Values
4. SELECT count(\*) from `scaler-target-project.products.data` where product\_photos\_qty IS NULL 610 0 Null Values
5. SELECT count(\*) from `scaler-target-project.products.data` where product\_weight\_g IS NULL 2 0 Null Values
6. SELECT count(\*) from `scaler-target-project.products.data` where product\_length\_cm IS NULL 2 0 Null Values
7. SELECT count(\*) from `scaler-target-project.products.data` where product\_height\_cm IS NULL 2 0 Null Values
8. SELECT count(\*) from `scaler-target-project.products.data` where product\_width\_cm IS NULL 2 0 Null Values

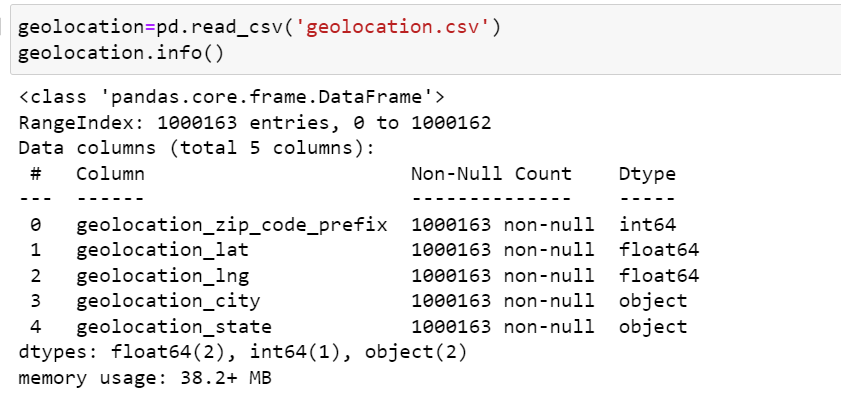
For Geolocation Table :

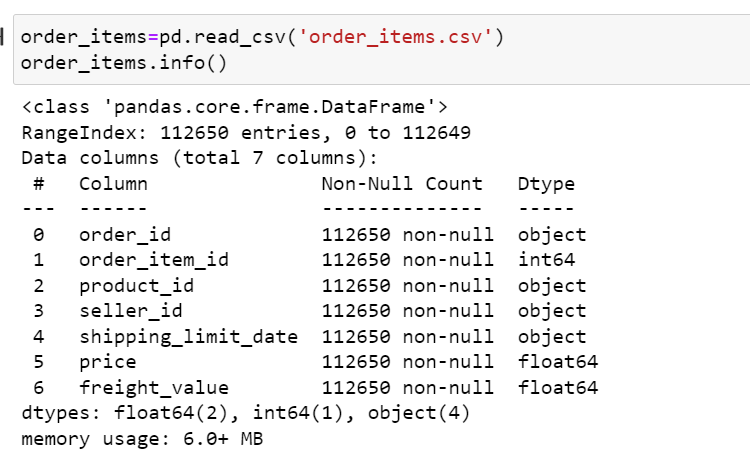
1. SELECT count(\*) from `scaler-target-project.geolocation.data` where geolocation\_zip\_code\_prefix IS NULL 0 Null Values
2. SELECT count(\*) from `scaler-target-project.geolocation.data` where geolocation\_lat IS NULL 0 Null Values
3. SELECT count(\*) from `scaler-target-project.geolocation.data` where geolocation\_lng IS NULL 0 Null Values
4. SELECT count(\*) from `scaler-target-project.geolocation.data` where geolocation\_city IS NULL 0 Null Values
5. SELECT count(\*) from `scaler-target-project.geolocation.data` where geolocation\_state IS NULL 0 Null Values
   1. Data type of columns in a table

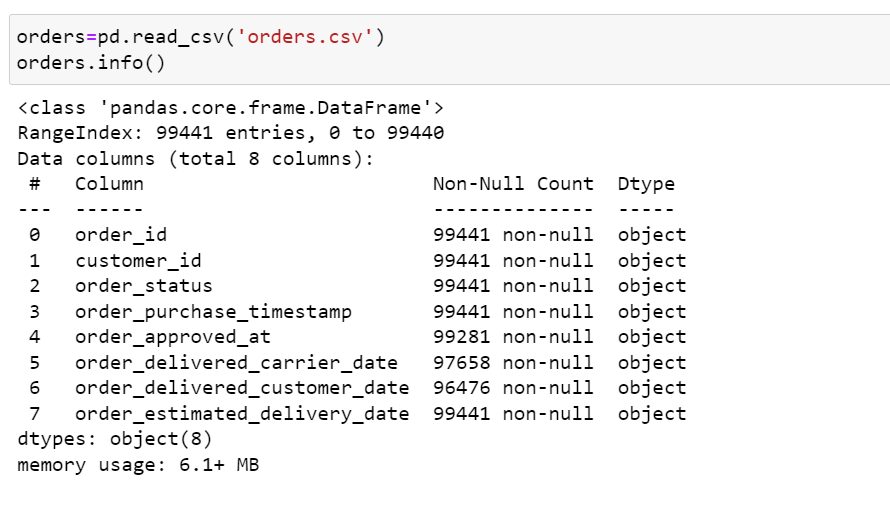
We can see the data type of columns of a particular table using DESC Table\_Name. This command is however not working in BigQuery. I have used Python to get the data types of the columns.



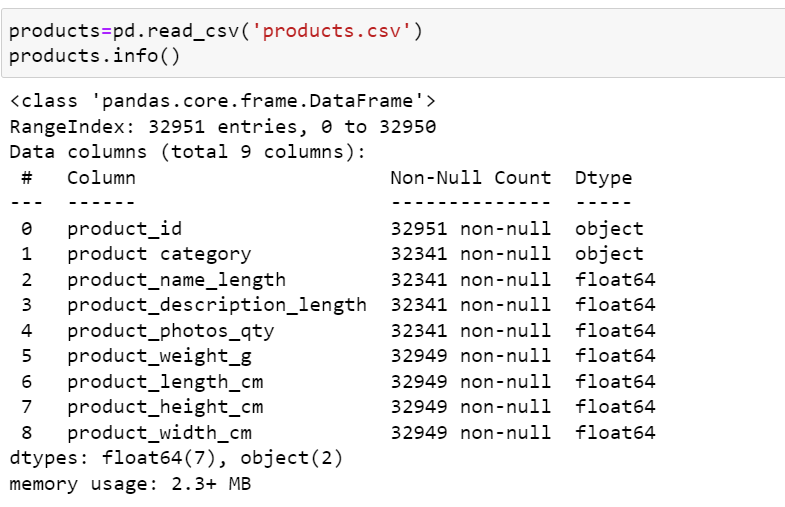


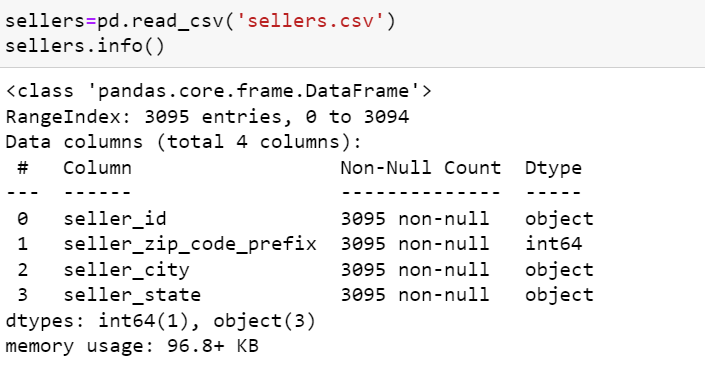






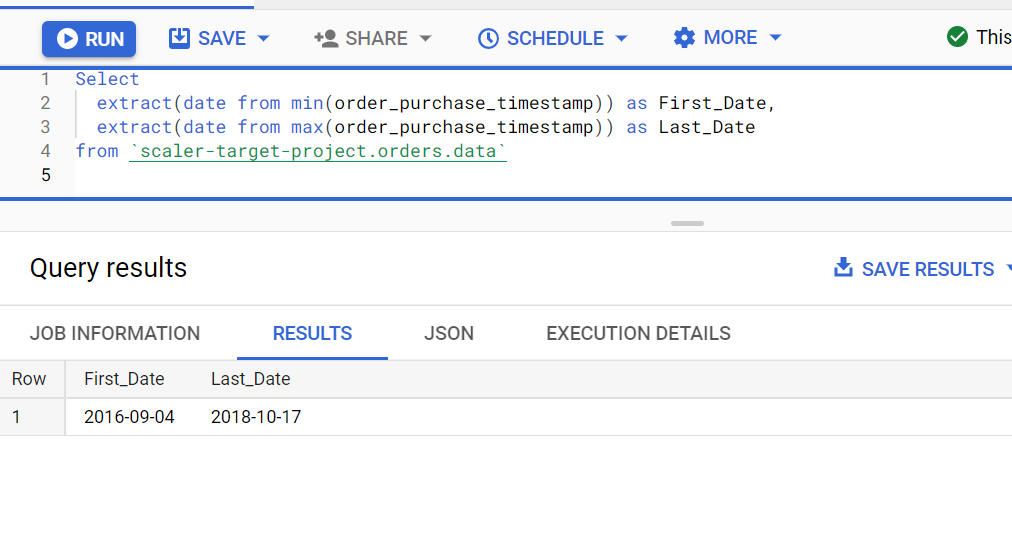






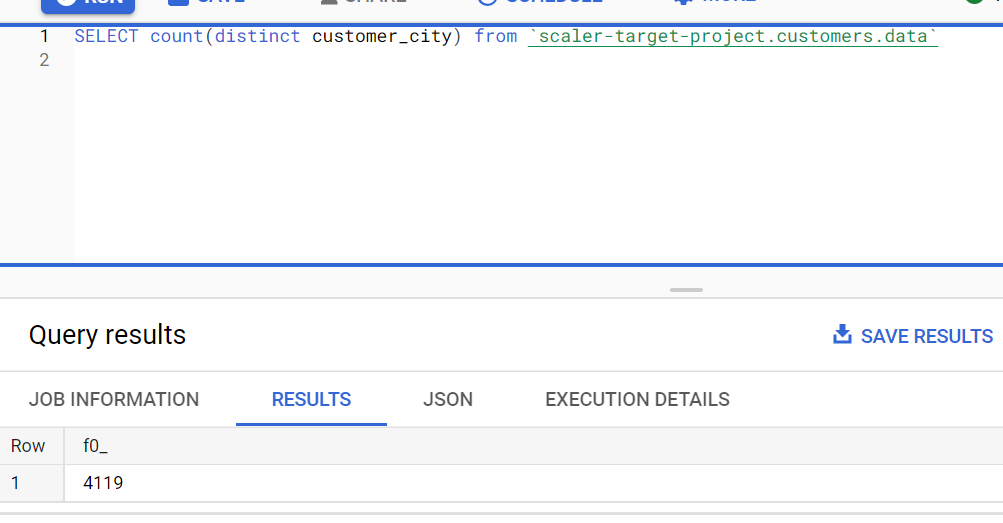
* 1. Get the time period for which the data is given

The time period of the data can be found from the order\_purchase\_timestamp column from orders table. We can find the first and the last order\_purchase\_timestamp which would give us the time period of the given data.

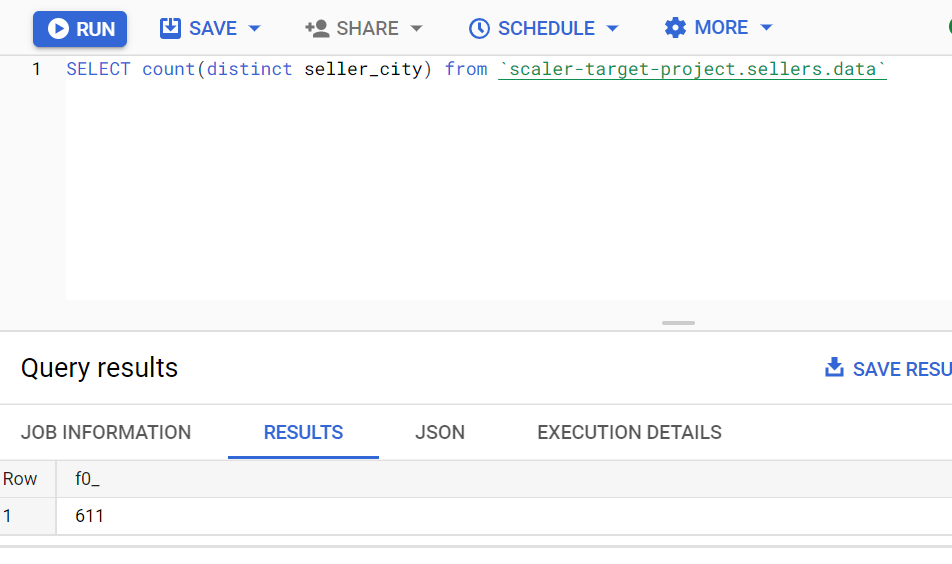


* 1. Number of cities in our dataset

No Of Customer\_Cities in Customer Table – 4119

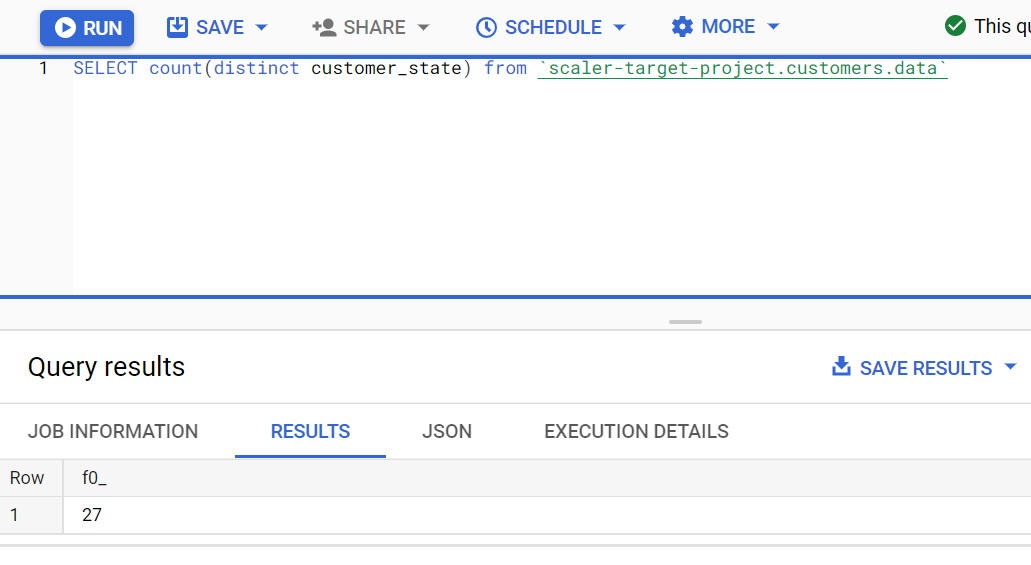


No Of Seller\_Cities in Seller Table – 611

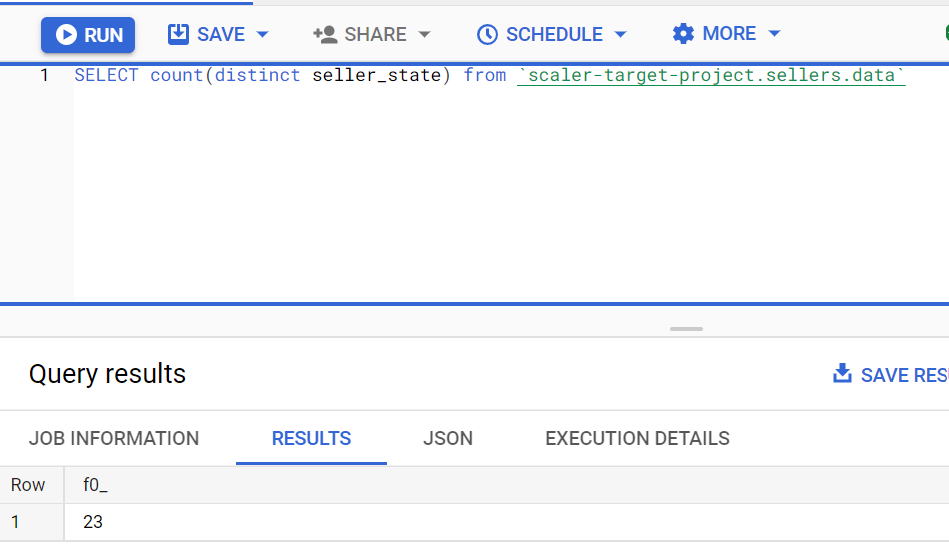


* 1. Number of states in our dataset

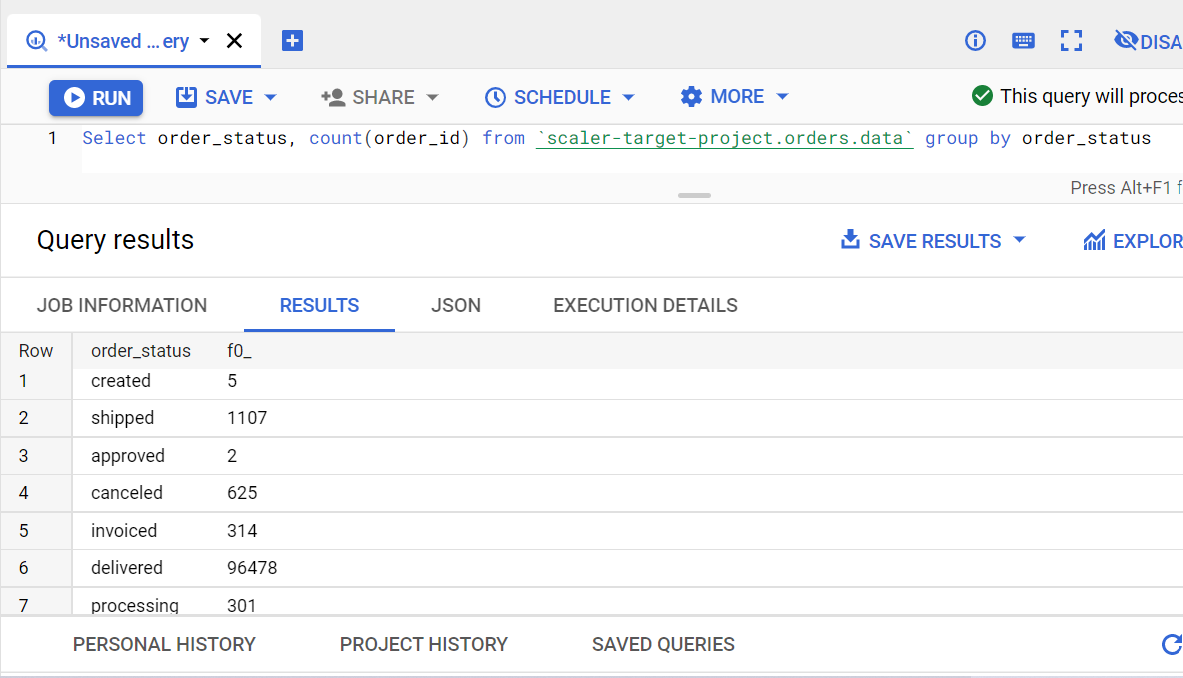
No Of Customer\_States in Customer Table – 27



No Of Seller\_States in Sellers Table - 23



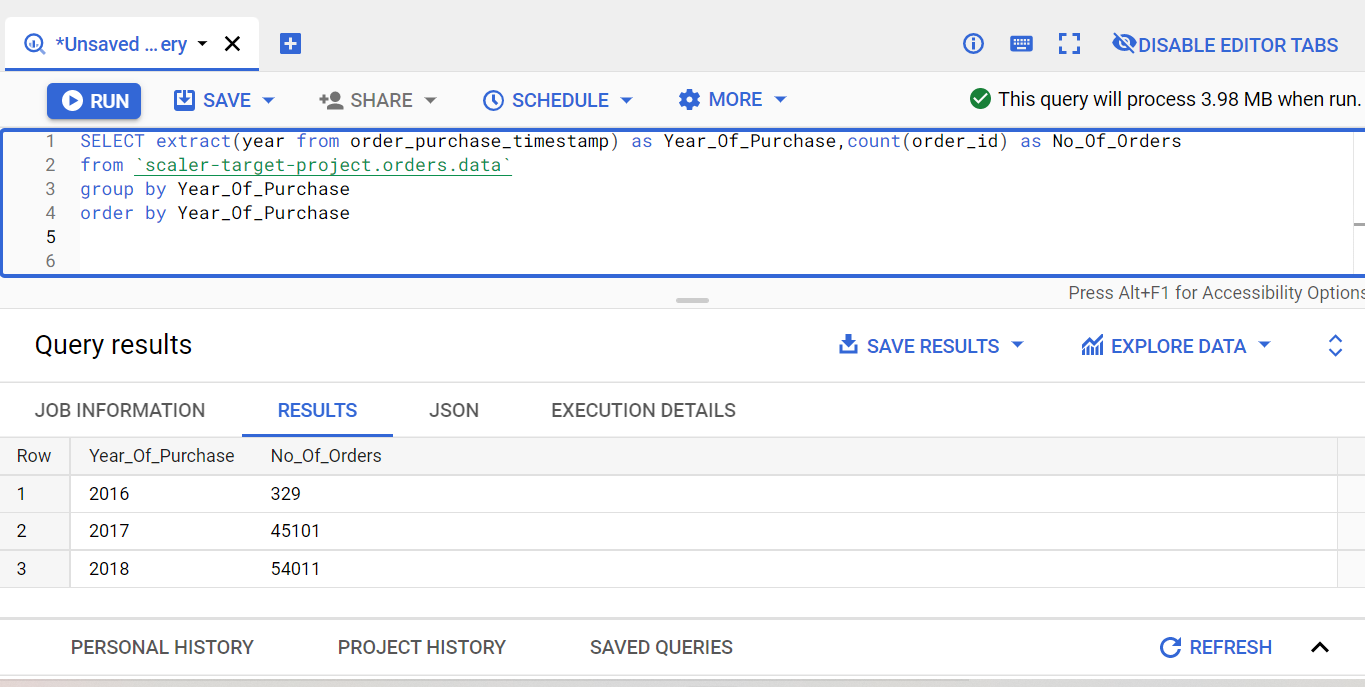
* In-depth Exploration:  
  1. How many orders do we have for each order status?



|  |  |  |  |
| --- | --- | --- | --- |
| 1 | created | 5 |  |
| 2 | shipped | 1107 |  |
| 3 | approved | 2 |  |
| 4 | canceled | 625 |  |
| 5 | invoiced | 314 |  |
| 6 | delivered | 96478 |  |
| 7 | processing | 301 |  |
| 8 | unavailable | 609 |  |

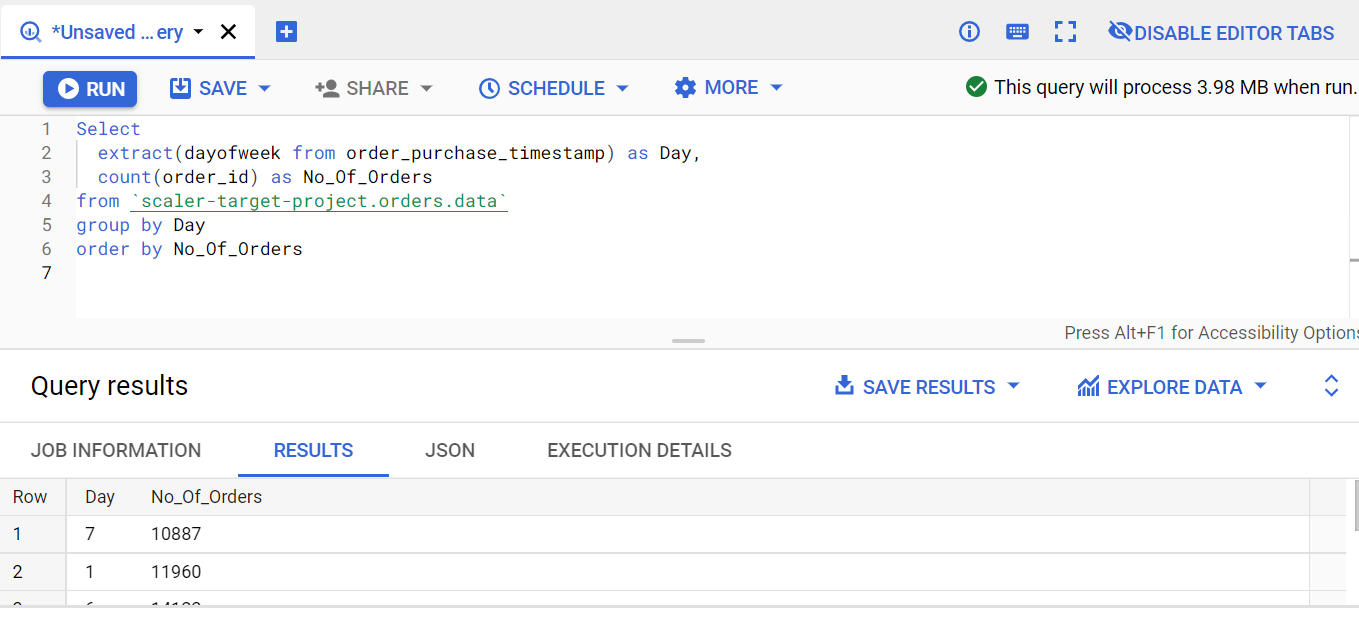
* 1. Is there a growing trend on e-commerce in Brazil? How can we describe a complete scenario?

We can confirm from the results that there was a growing trend in Brazil as there is Year on Year increase in the number of orders.



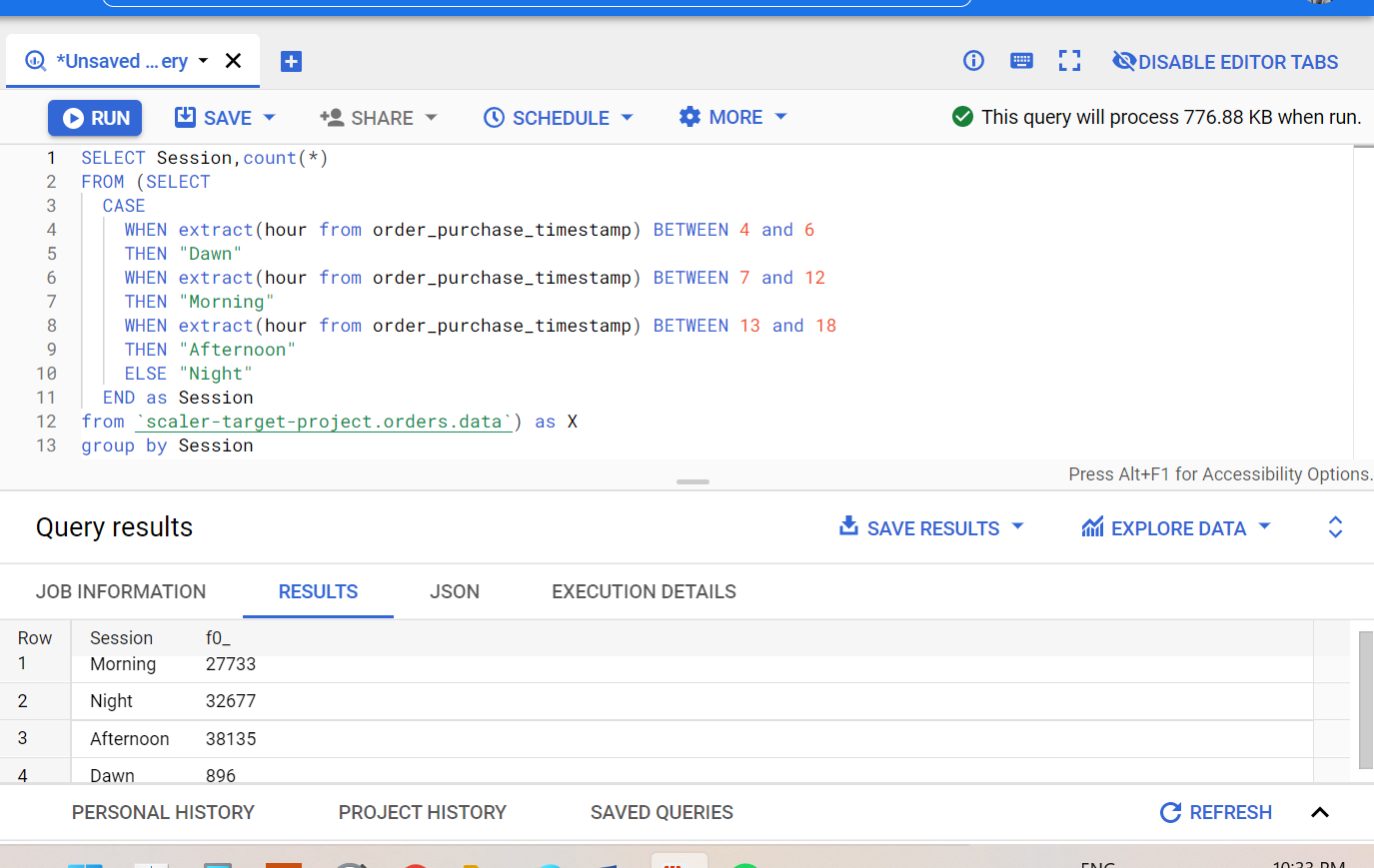
* 1. On what day of week brazilians customers tend to do online purchasing?

Assuming Day 1 to be Monday, maximum orders are placed on Day 3 which is Wednesday.

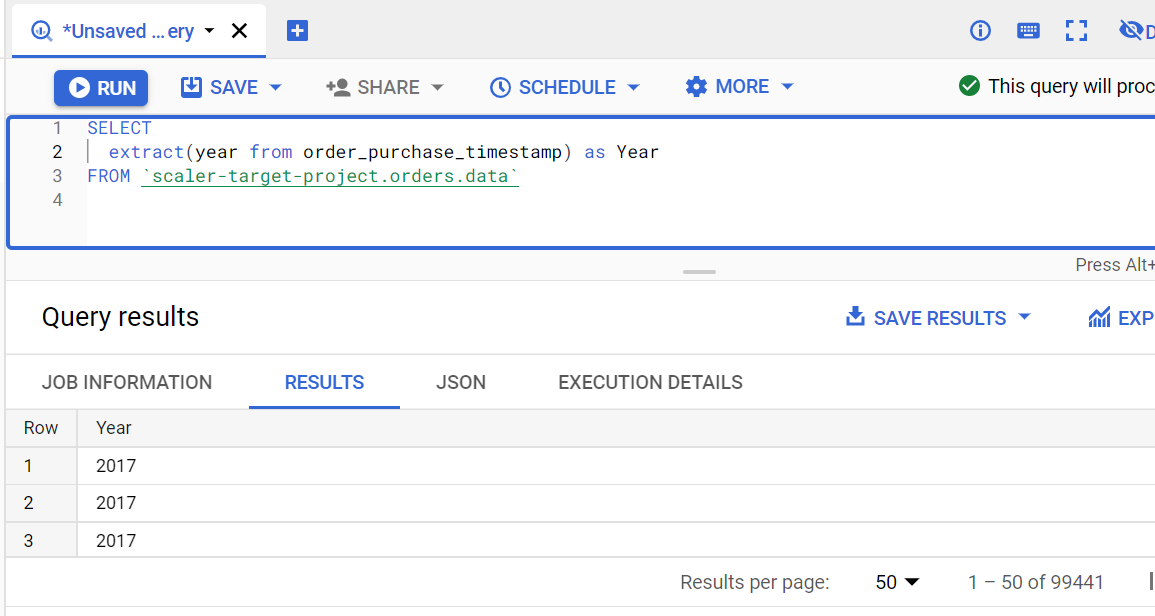


| **Row** | **Day** | **No\_Of\_Orders** |  |
| --- | --- | --- | --- |
| 1 | 7 | 10887 |  |
| 2 | 1 | 11960 |  |
| 3 | 6 | 14122 |  |
| 4 | 5 | 14761 |  |
| 5 | 4 | 15552 |  |
| 6 | 3 | 15963 |  |

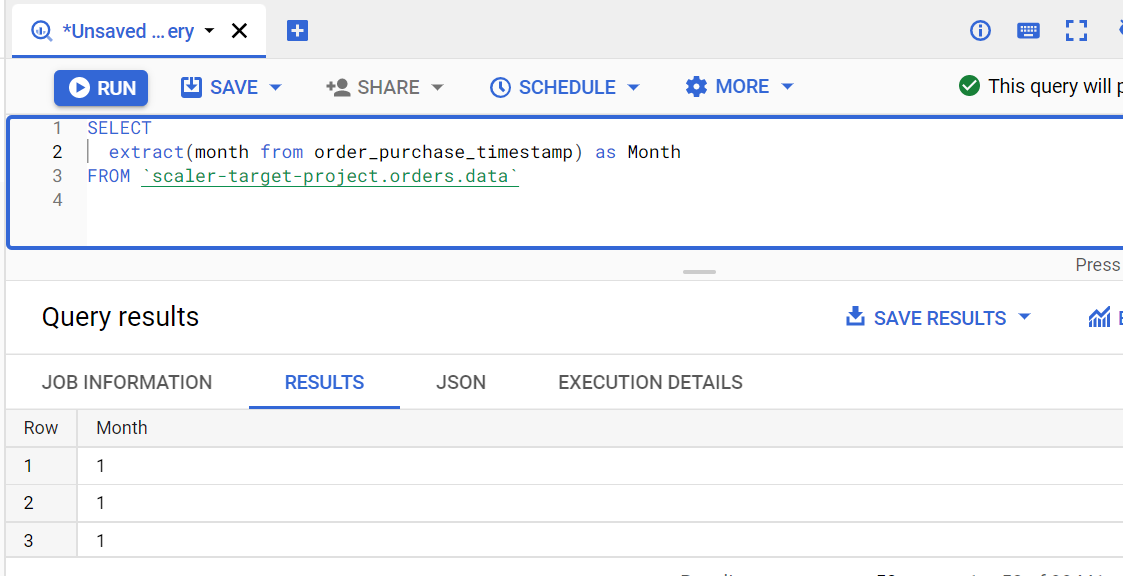
* 1. What time do Brazilian customers tend to buy (Dawn, Morning, Afternoon or Night)?



* 1. Feature Extraction: Through order\_purchase\_timestamp in “orders” dataset extract  
     1. order\_purchase\_year



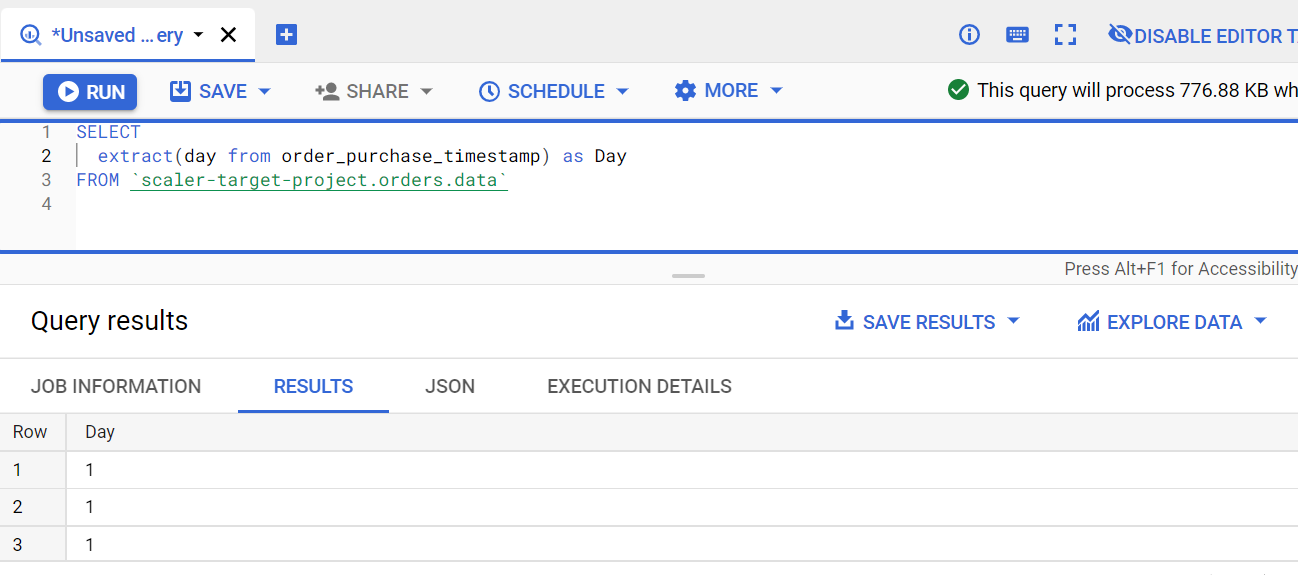
* + 1. order\_purchase\_month



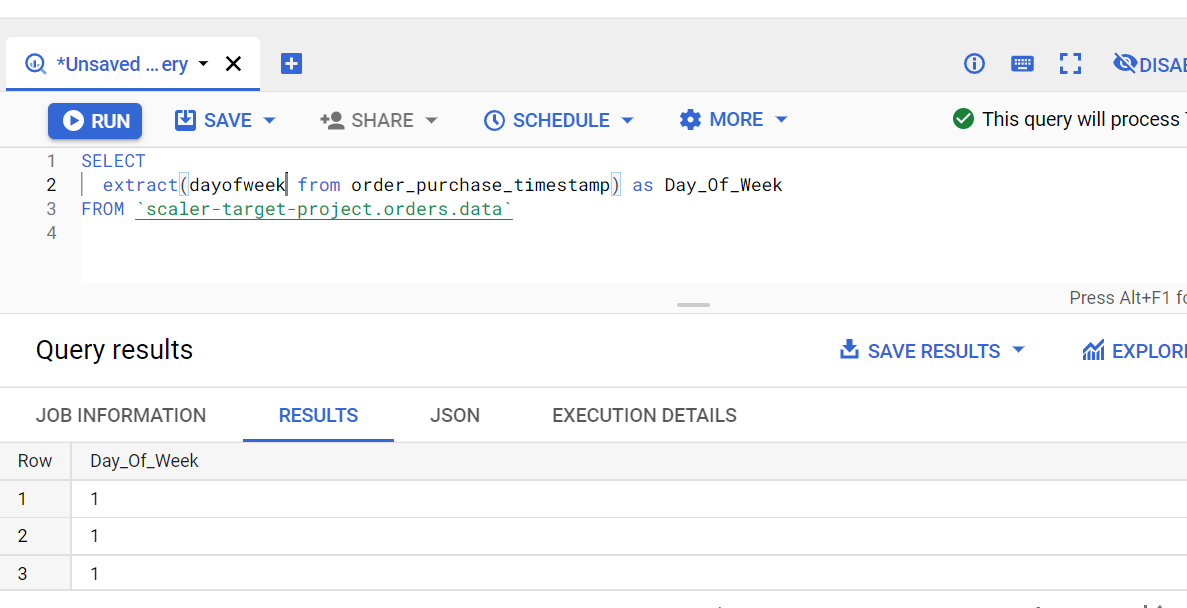
* + 1. order\_purchase\_date



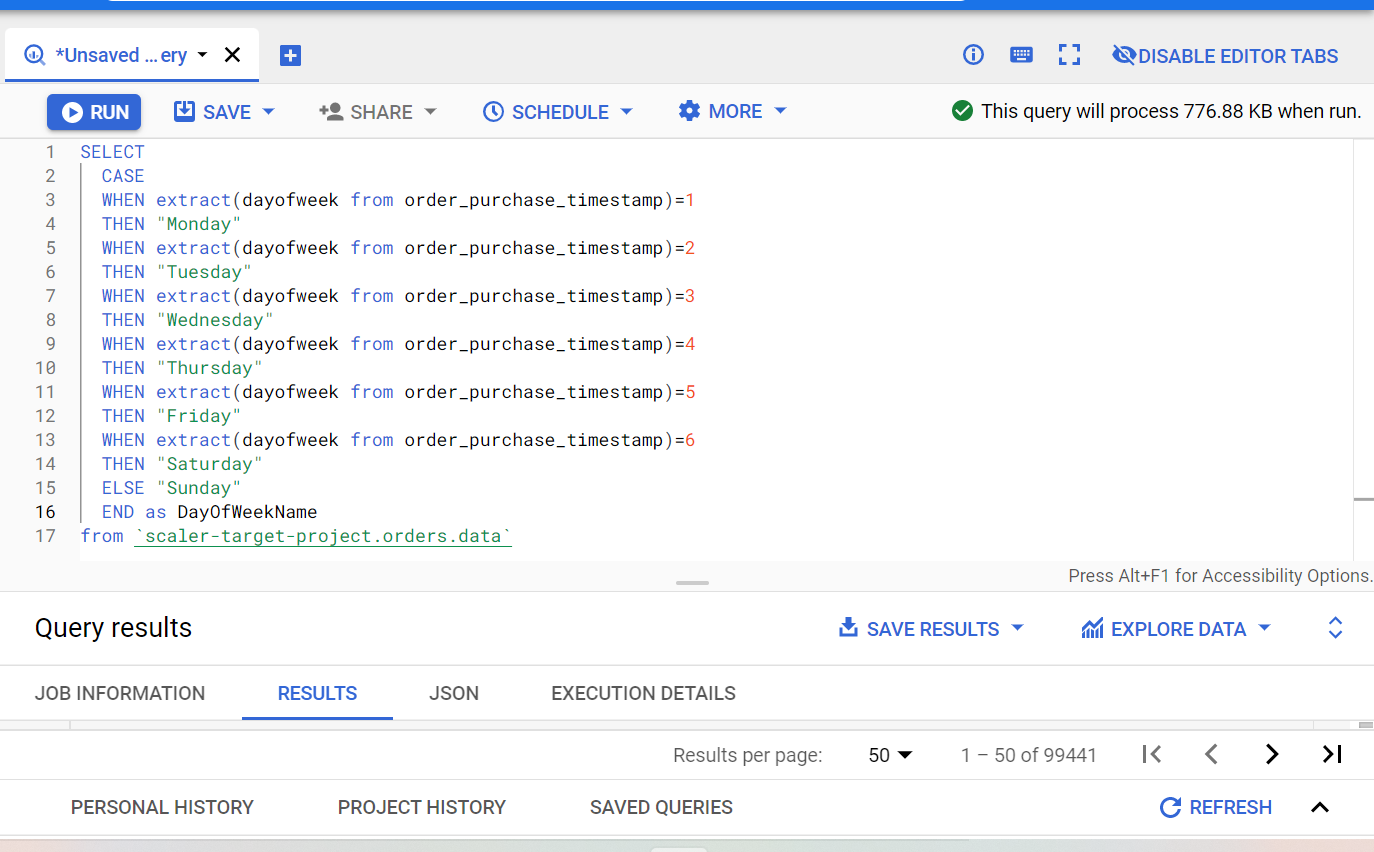
* + 1. order\_purchase\_day

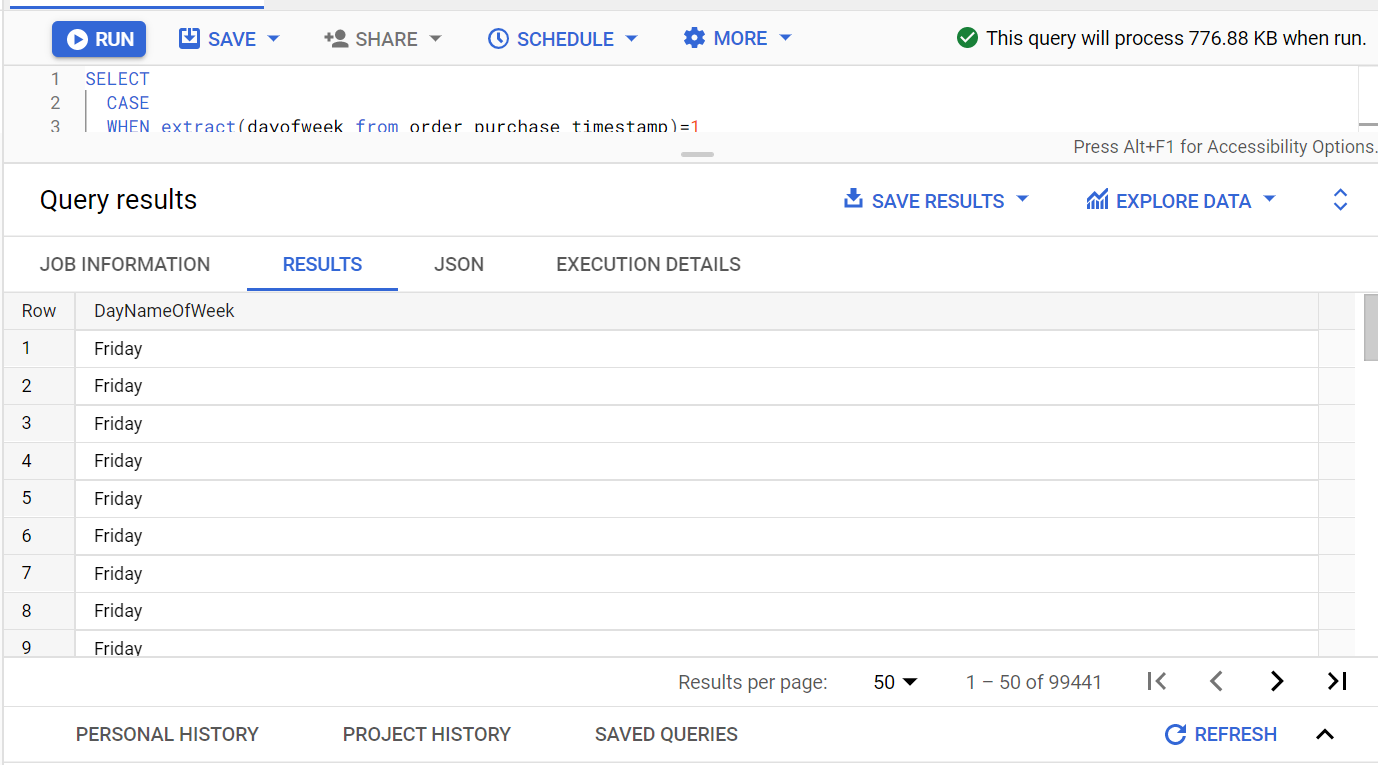


* + 1. order\_purchase\_dayofweek

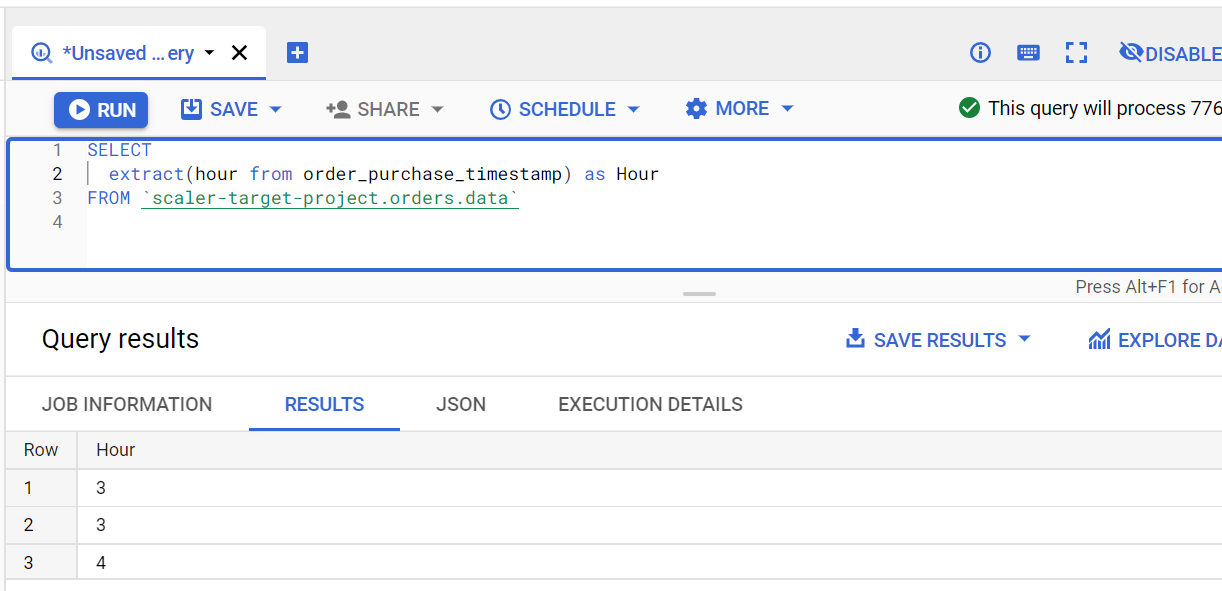


* + 1. order\_purchase\_dayofweek\_name

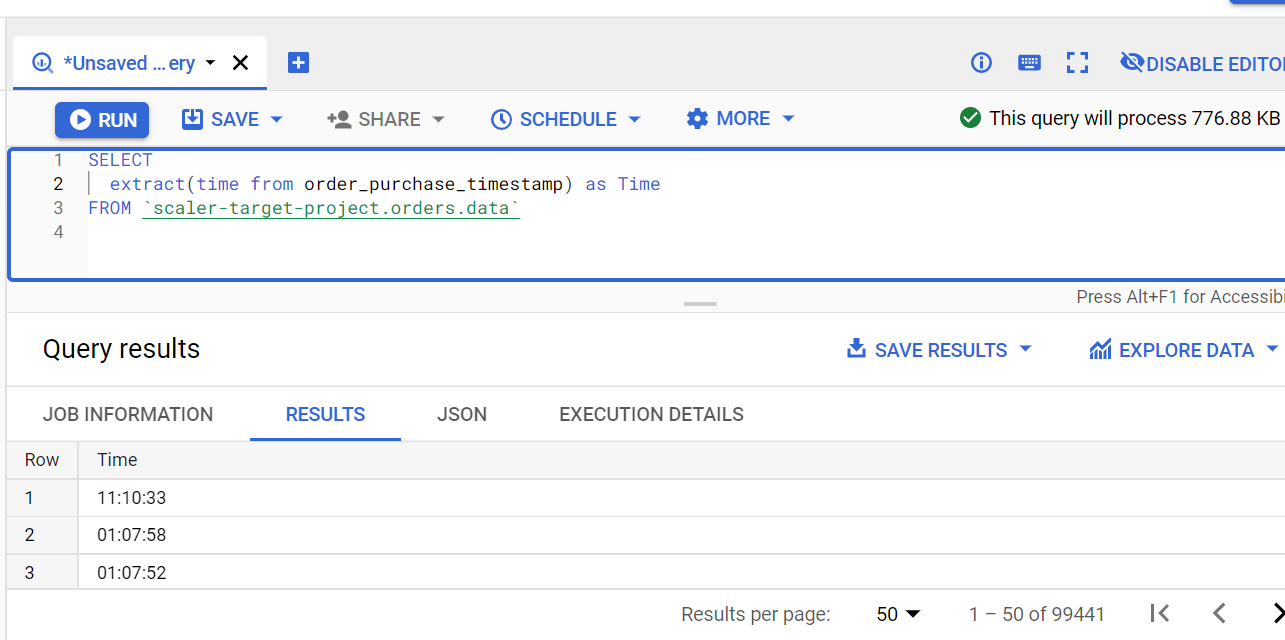




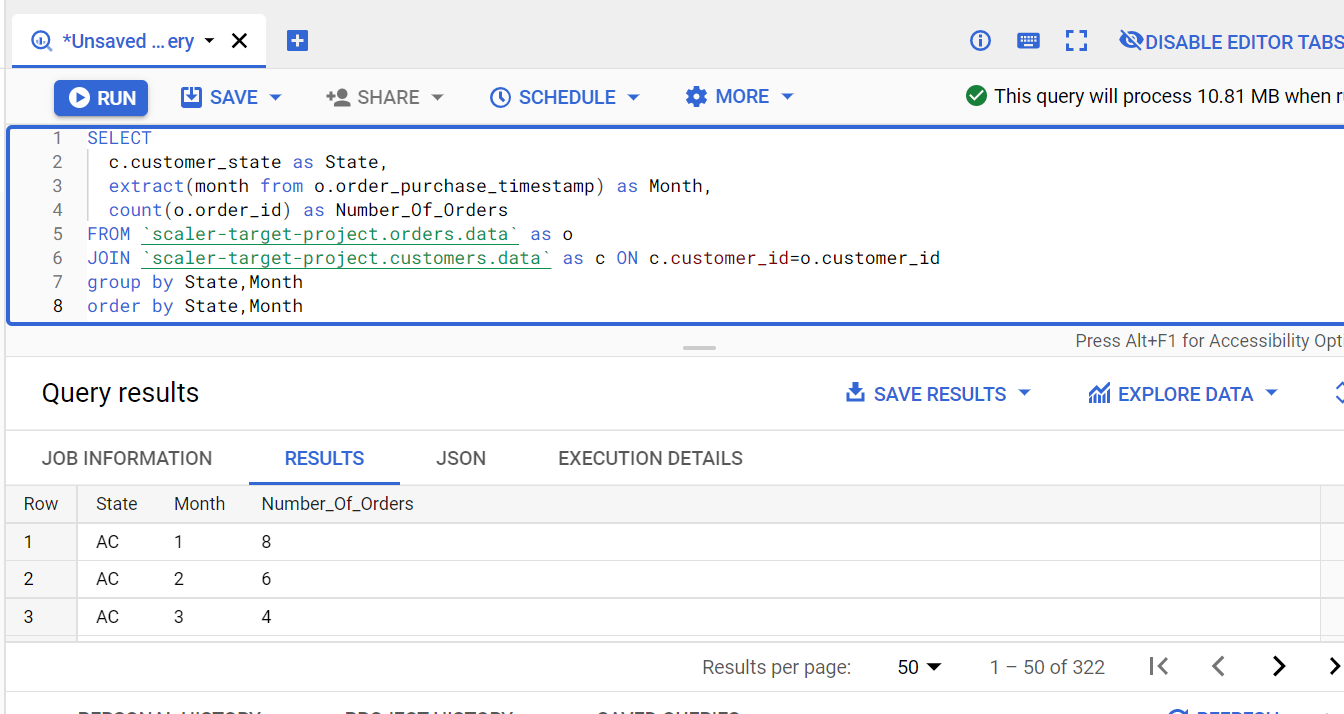
* + 1. order\_purchase\_hour



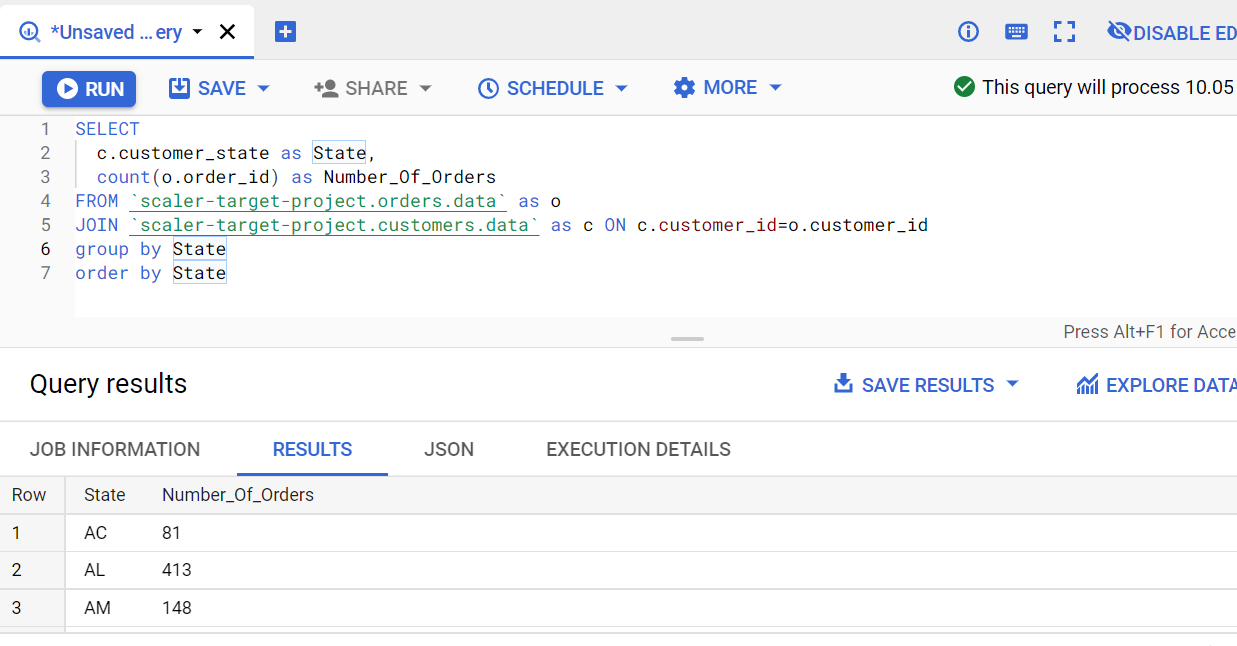
* + 1. order\_purchase\_time\_day



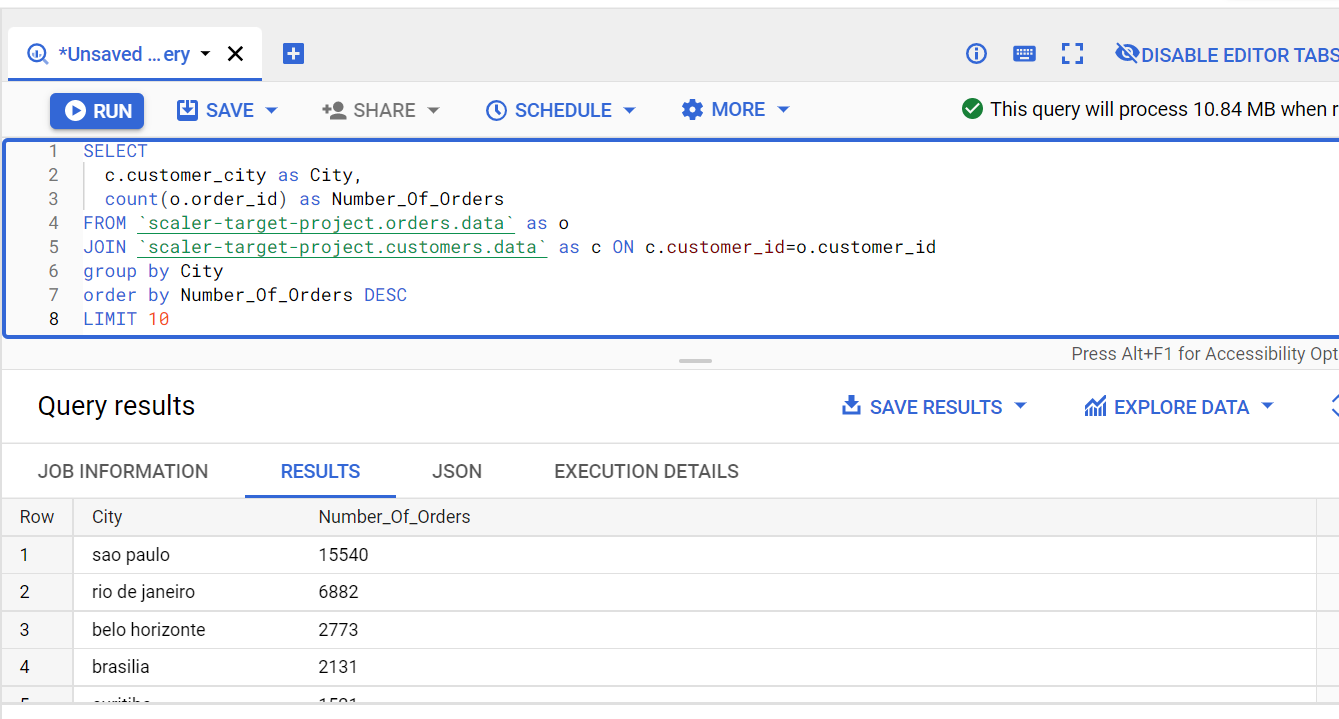
* Evolution of E-commerce orders in the Brazil region:  
  1. Get month on month orders by region



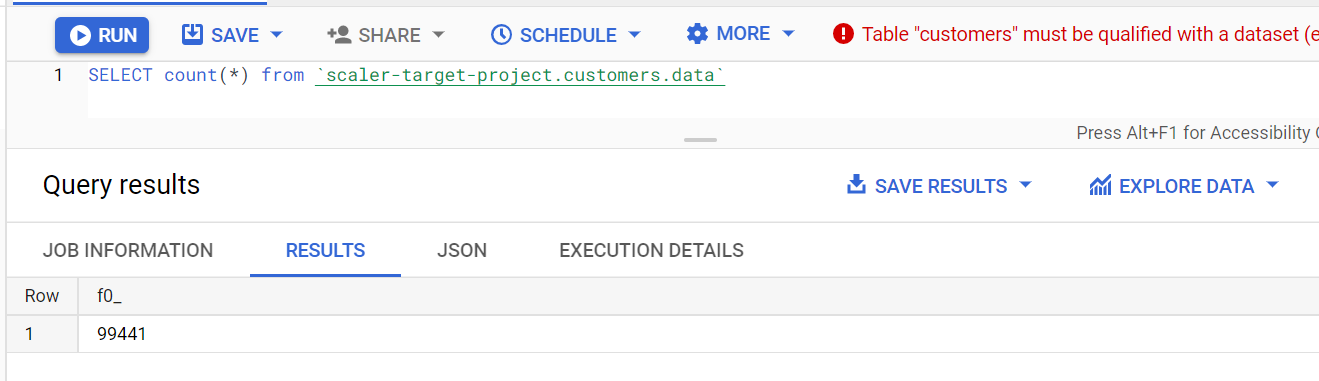
* 1. Total of customer orders by state

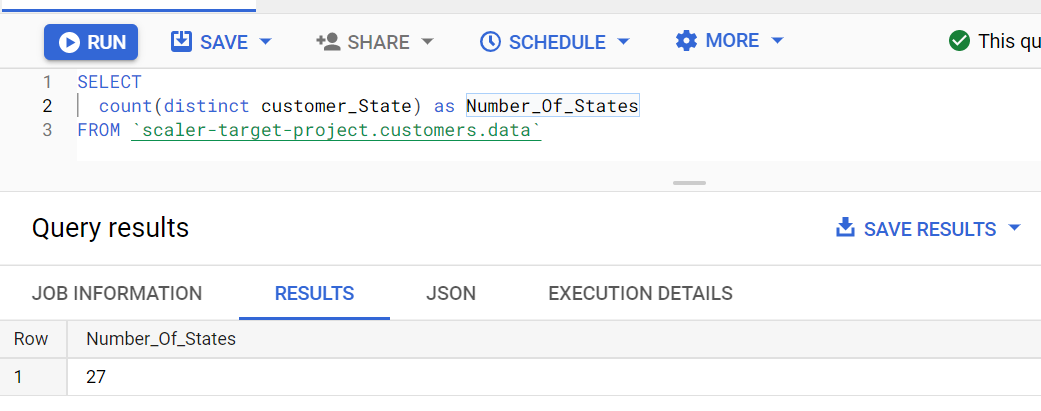


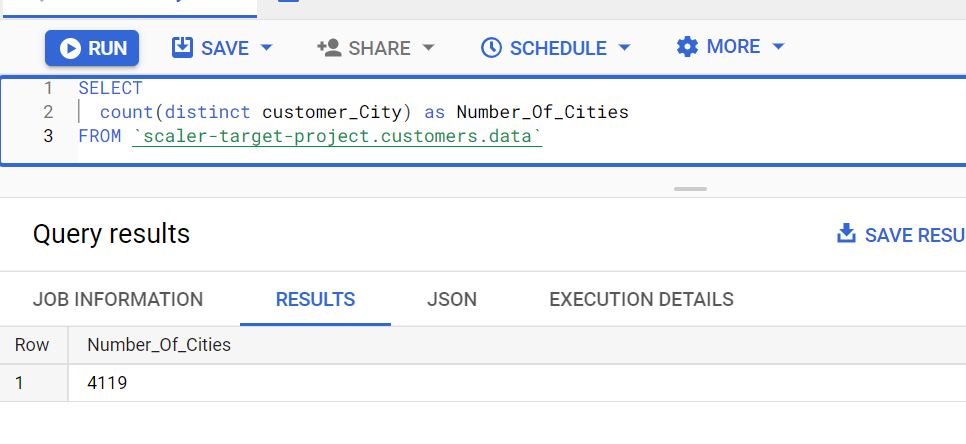
* 1. Top 10 brazilian cities most no. of orders

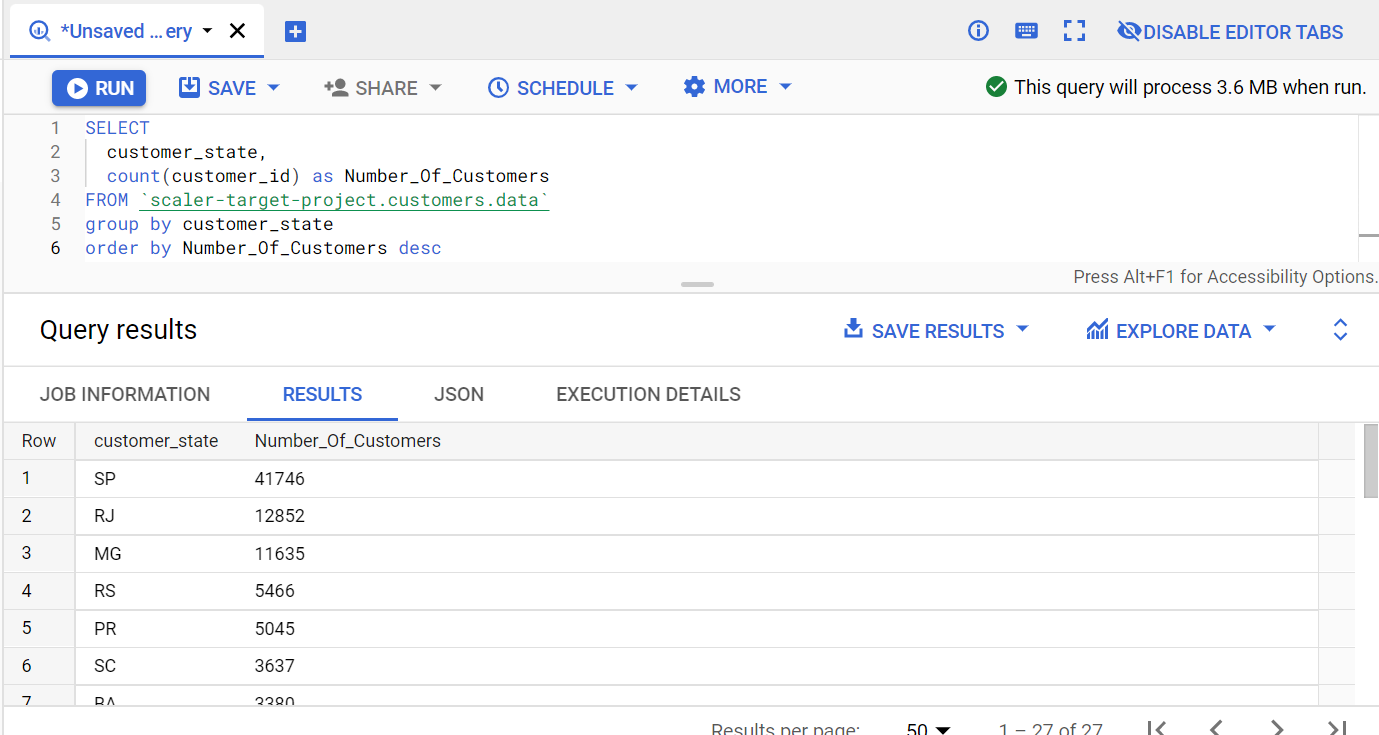


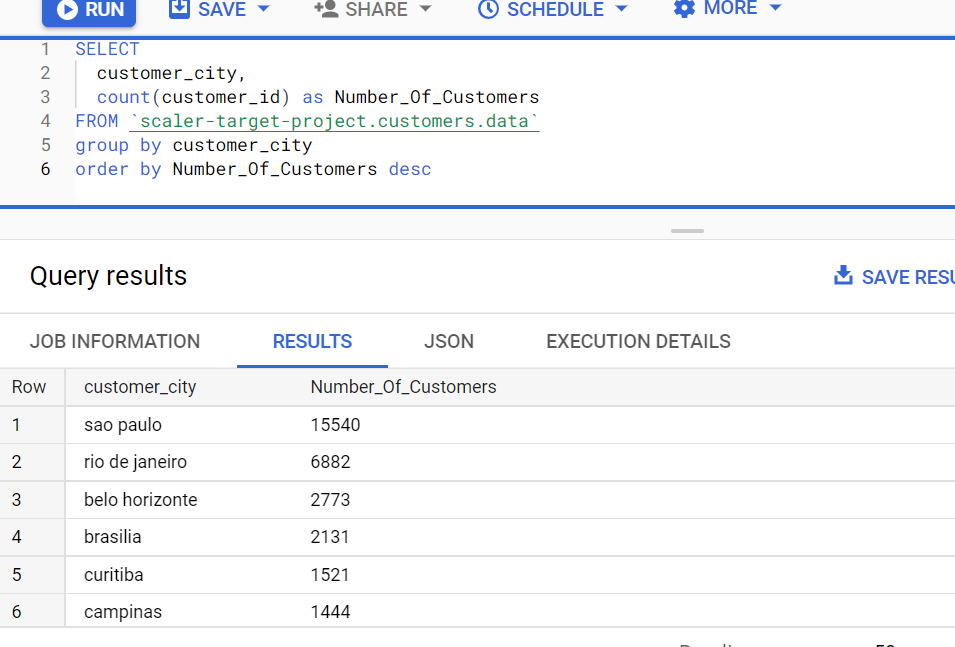
* 1. How are customers distributed in Brazil



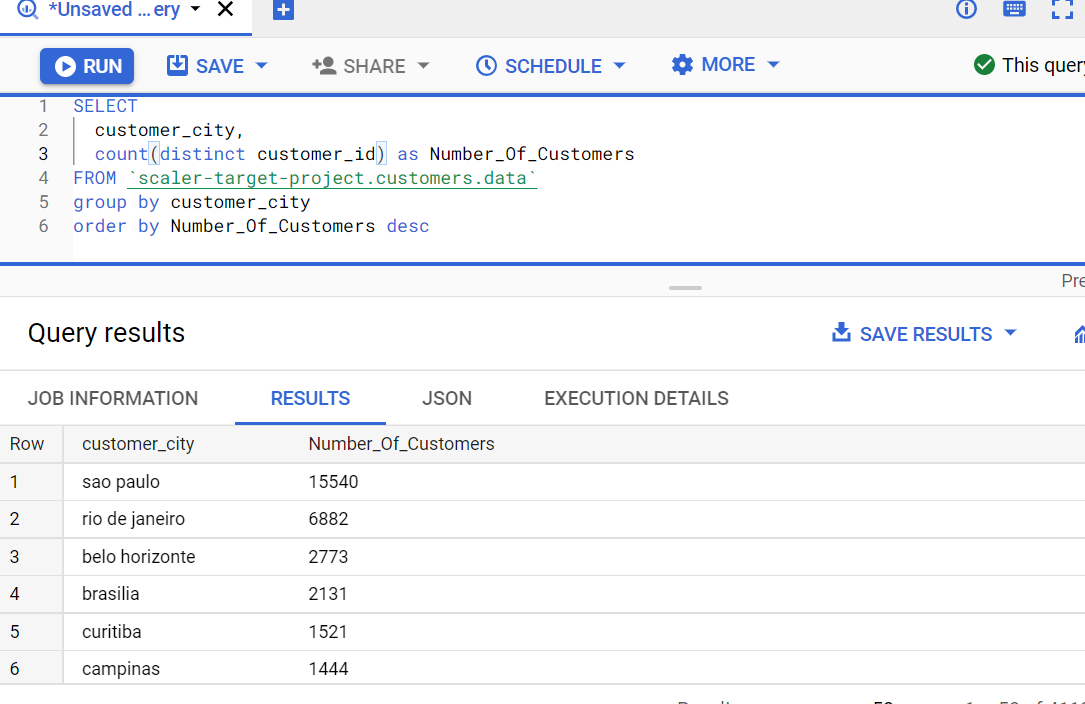








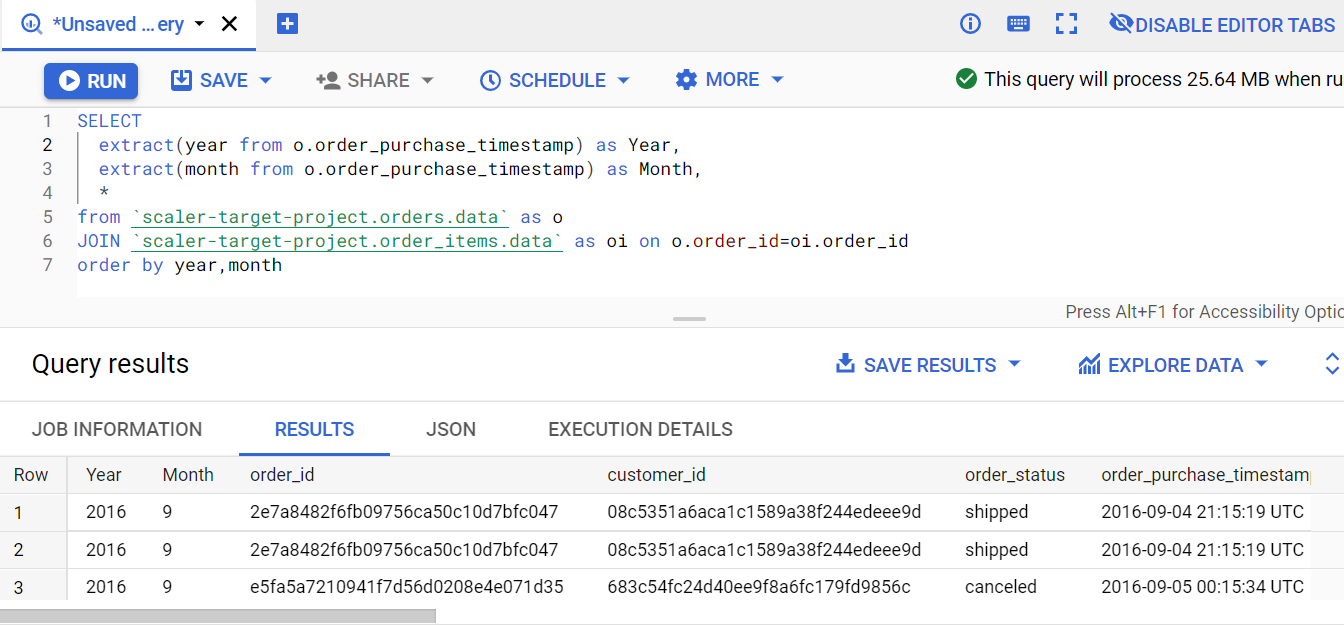
* 1. City wise number of unique customers



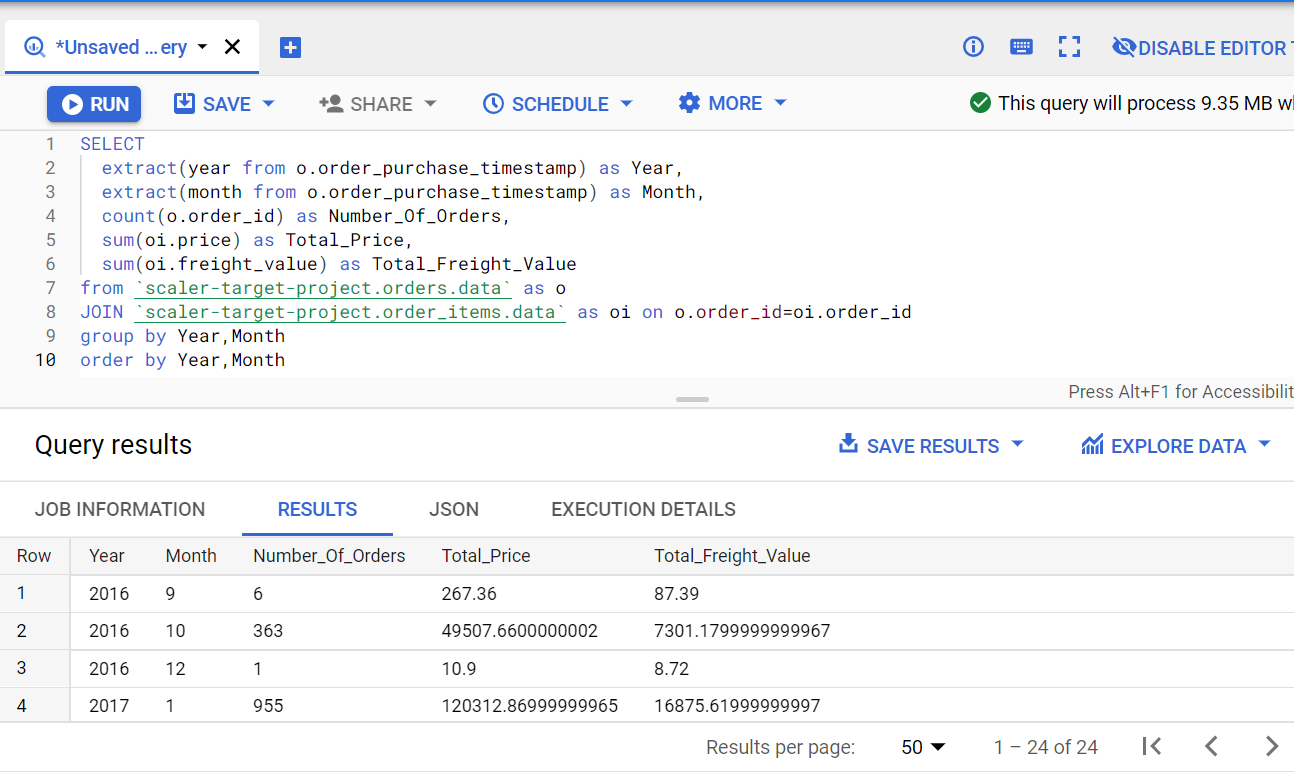
* Impact on Economy: Analyze the money movemented by e-commerce by looking at order prices, freight and others.

Step 1: Using CTE

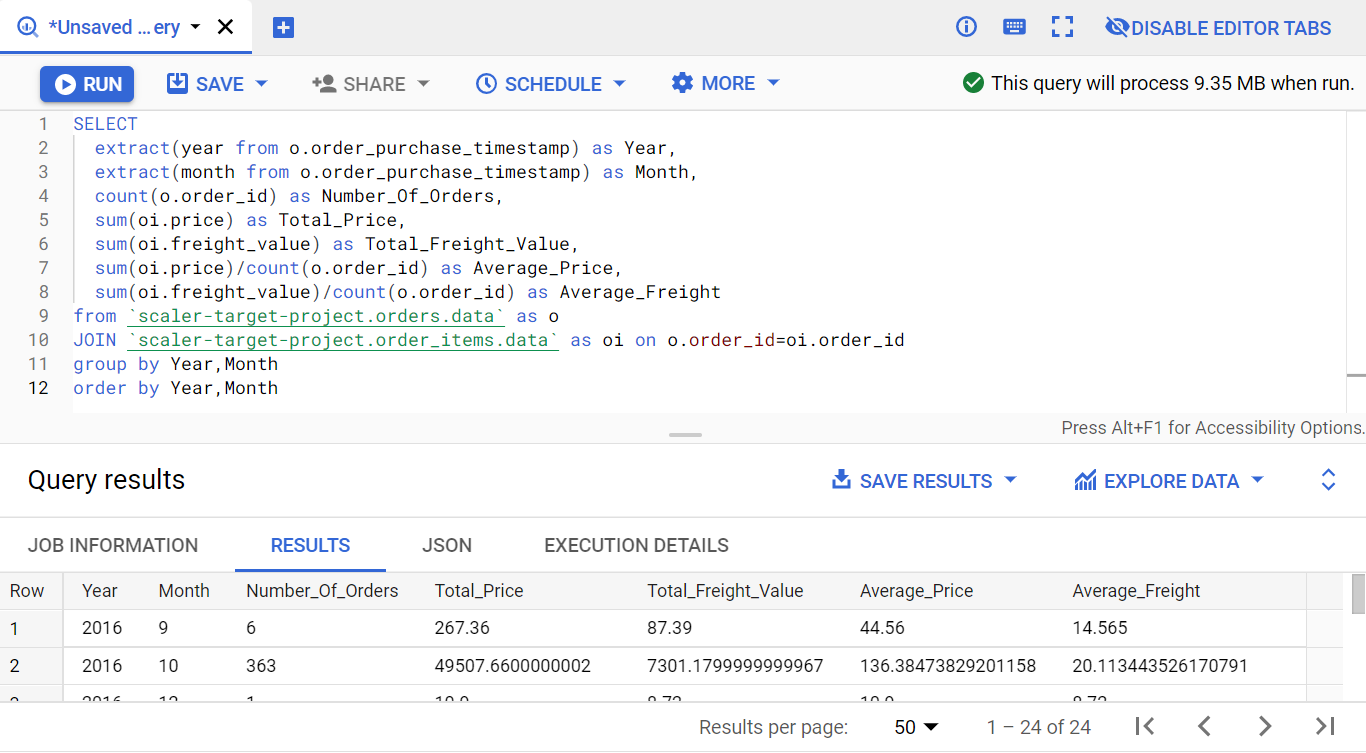
* 1. “order\_items” + “order” joined on order id where order\_purchase timestamp is already divided into month & year



2. Group data by year and month, aggregation count(order\_id), sum(price), sum(freight\_value)

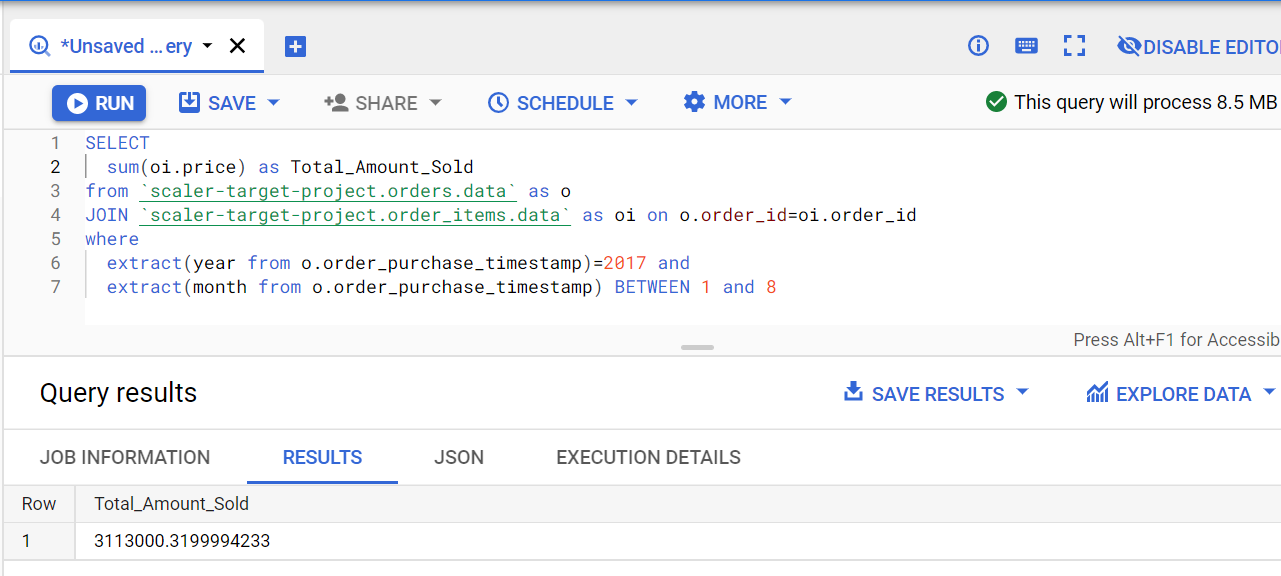


* 1. Create new columns:  
     price\_per\_order = sum(price) / count(order\_id)  
     freight\_per\_order= sum(freight\_value) / count(order\_id)

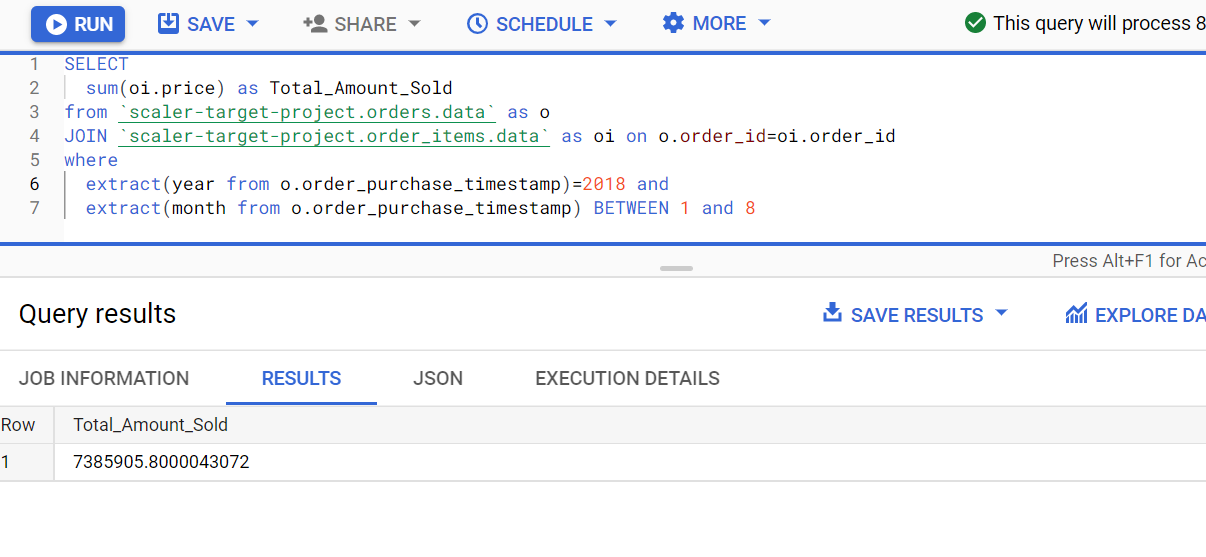


Step 2: Answer the following questions:

* + 1. Total amount sold in 2017 between Jan to August



2. Total amount sold in 2018 between Jan to august

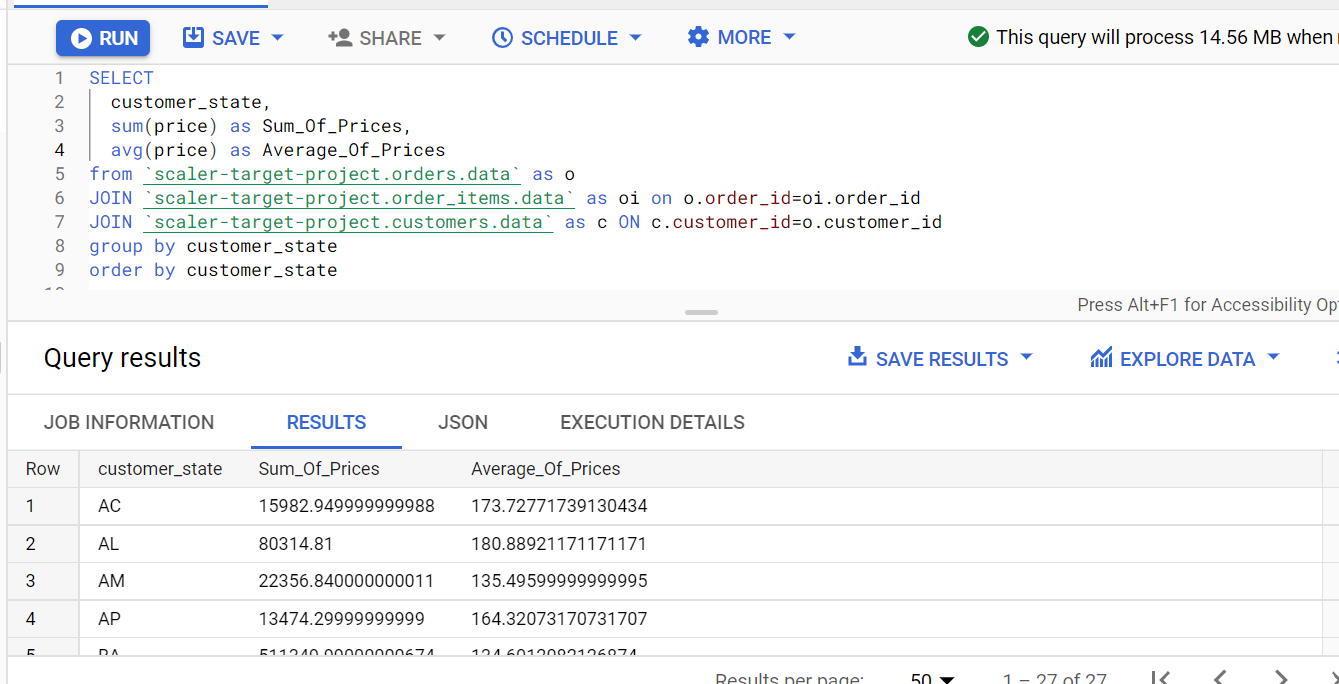


* 1. % increase from 2017 to 2018

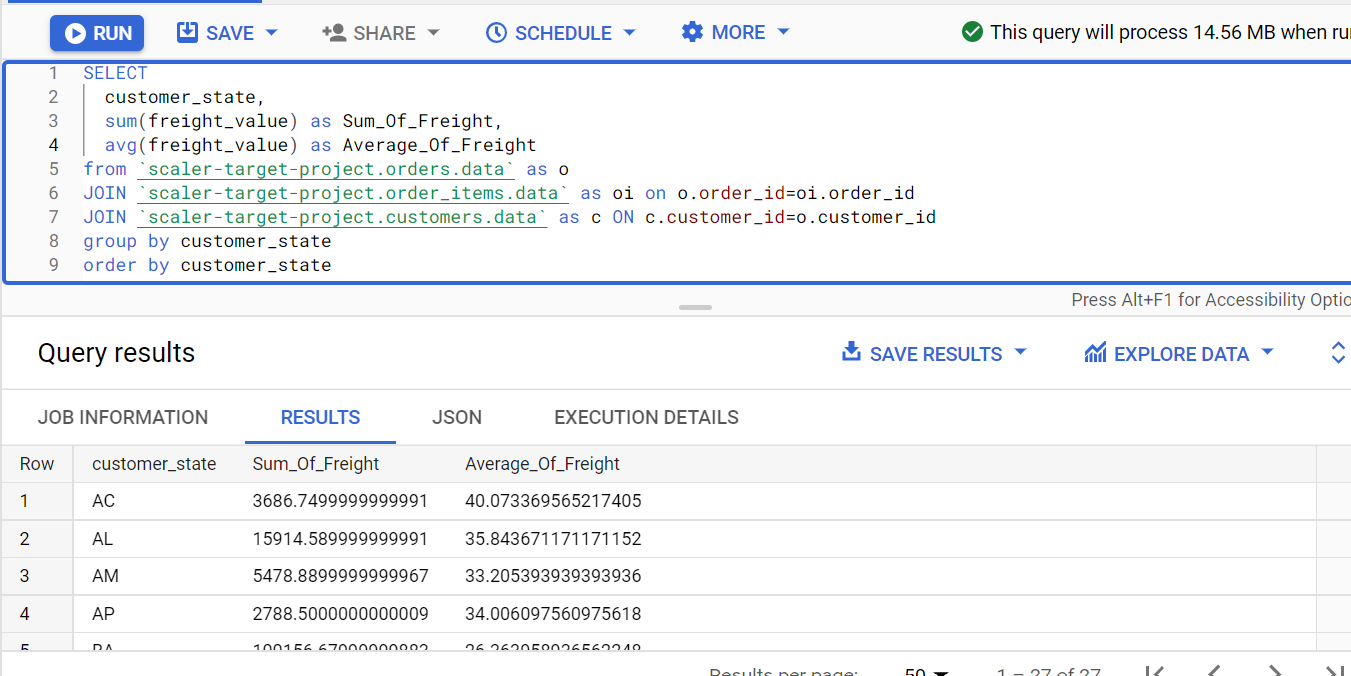


Step 3: Join (orders+order\_items) table from previous step with “customers” table on Customer\_id and find:

* + 1. Mean & Sum of price by customer state

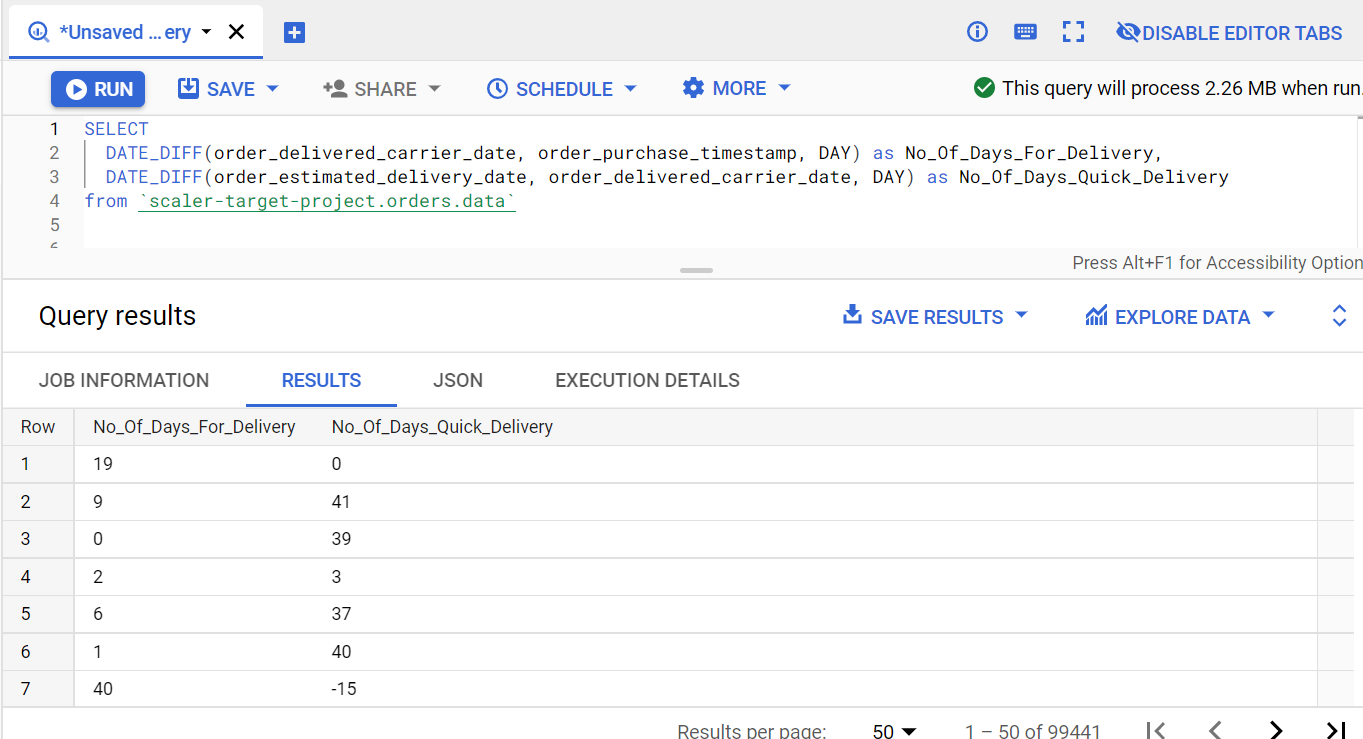


* + 1. Mean & Sum of freight value by customer state



5. Analysis on sales, freight and delivery time

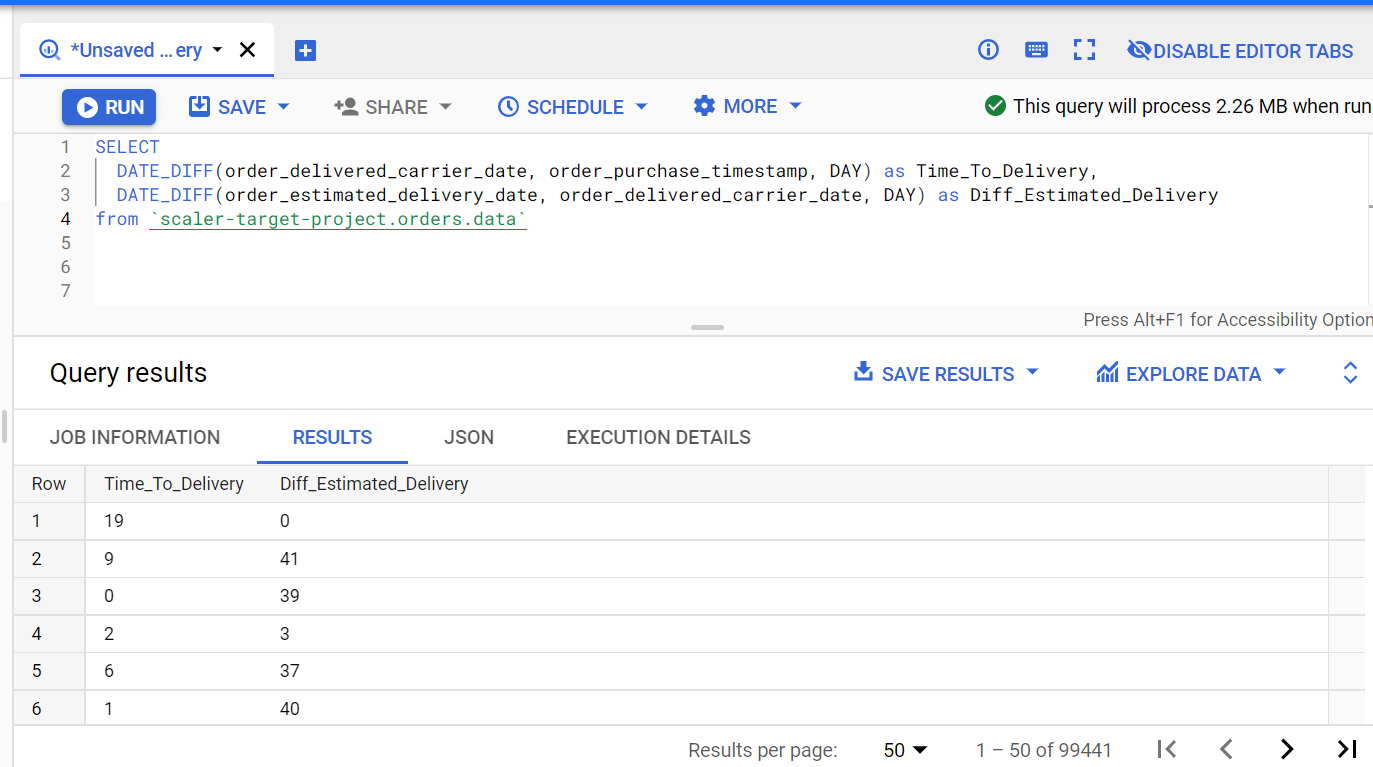
1. Calculating days between purchasing, delivering and estimated delivery



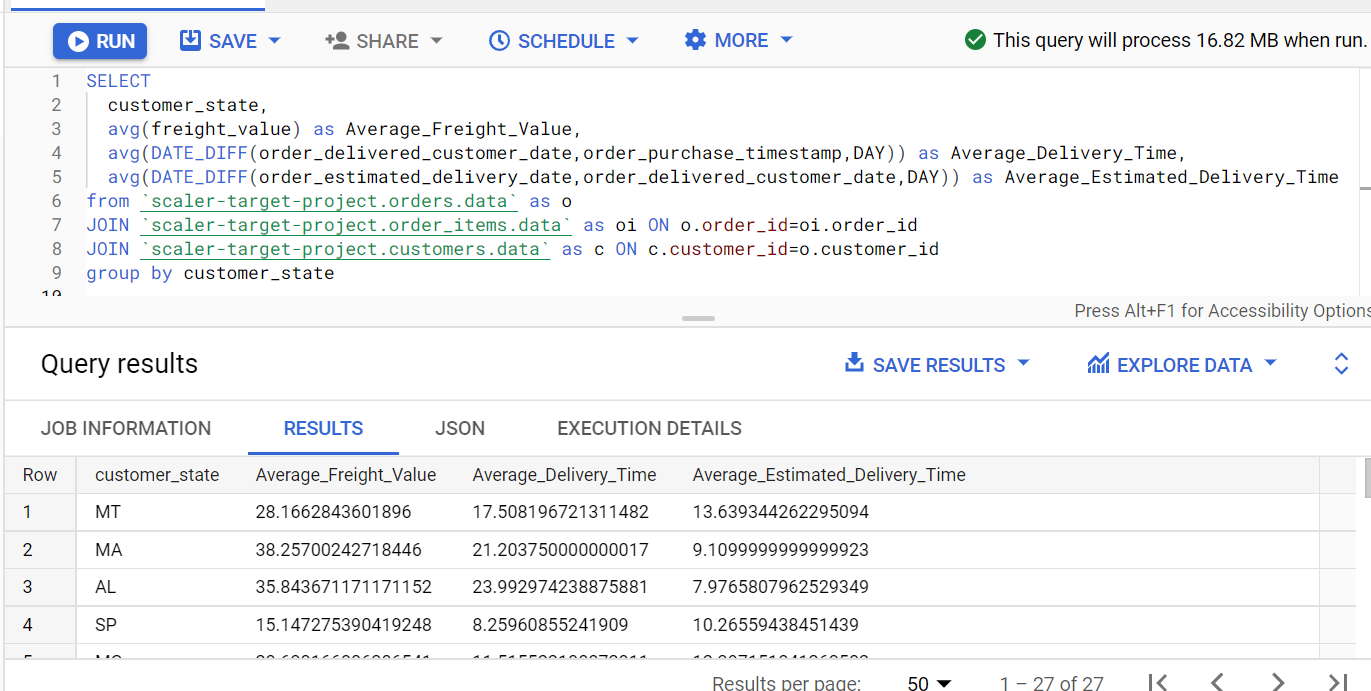
2. Create columns:

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

diff\_estimated\_delivery = order\_estimated\_delivery\_date-order\_delivered\_customer\_date



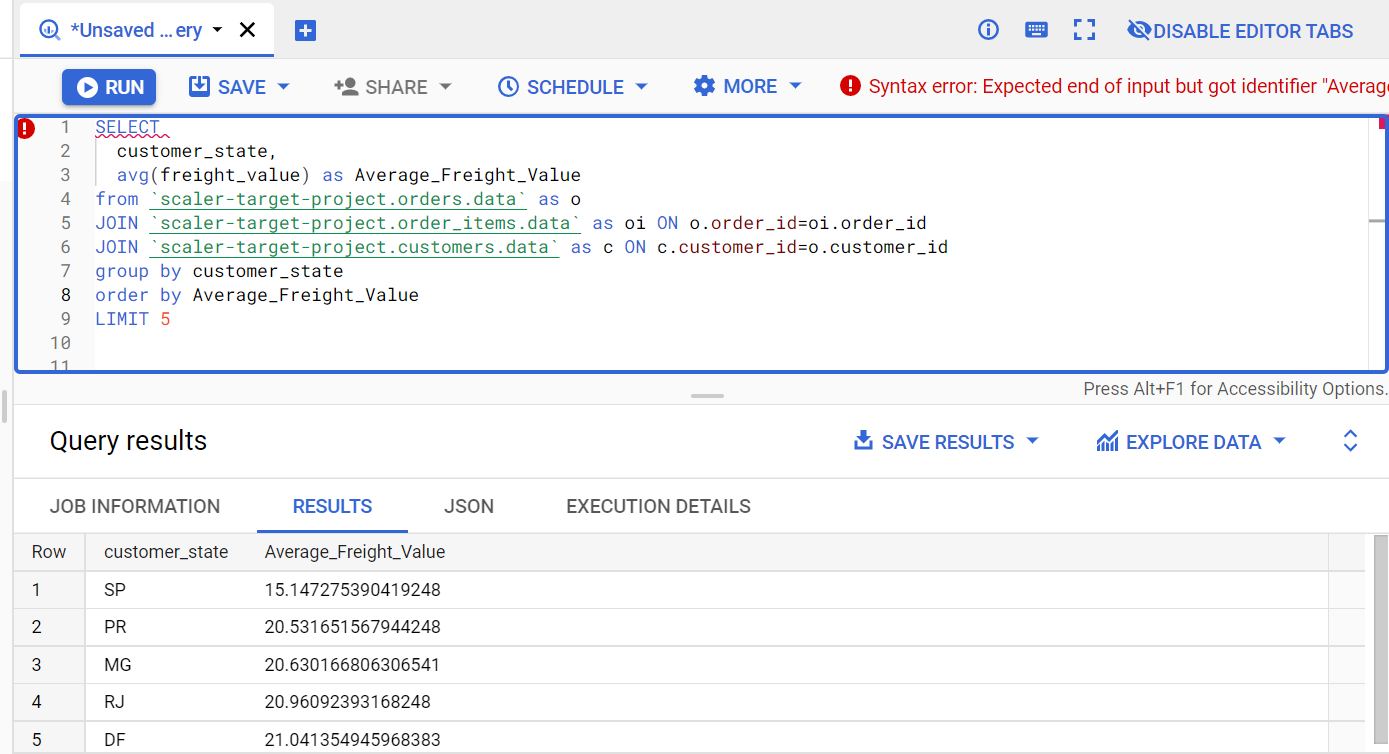
3. Grouping data by state, take mean of freight\_value, time\_to\_delivery, diff\_estimated\_delivery



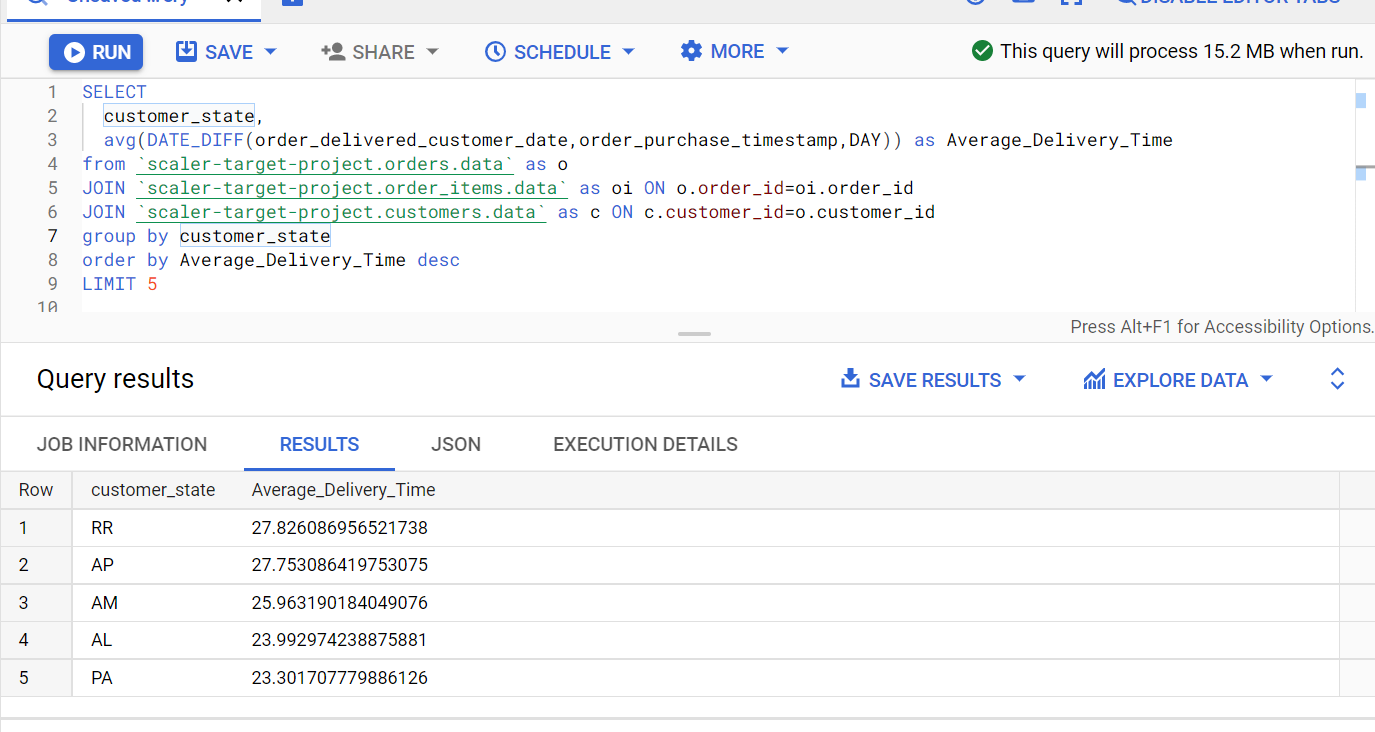
4. Sort the data to get the following:

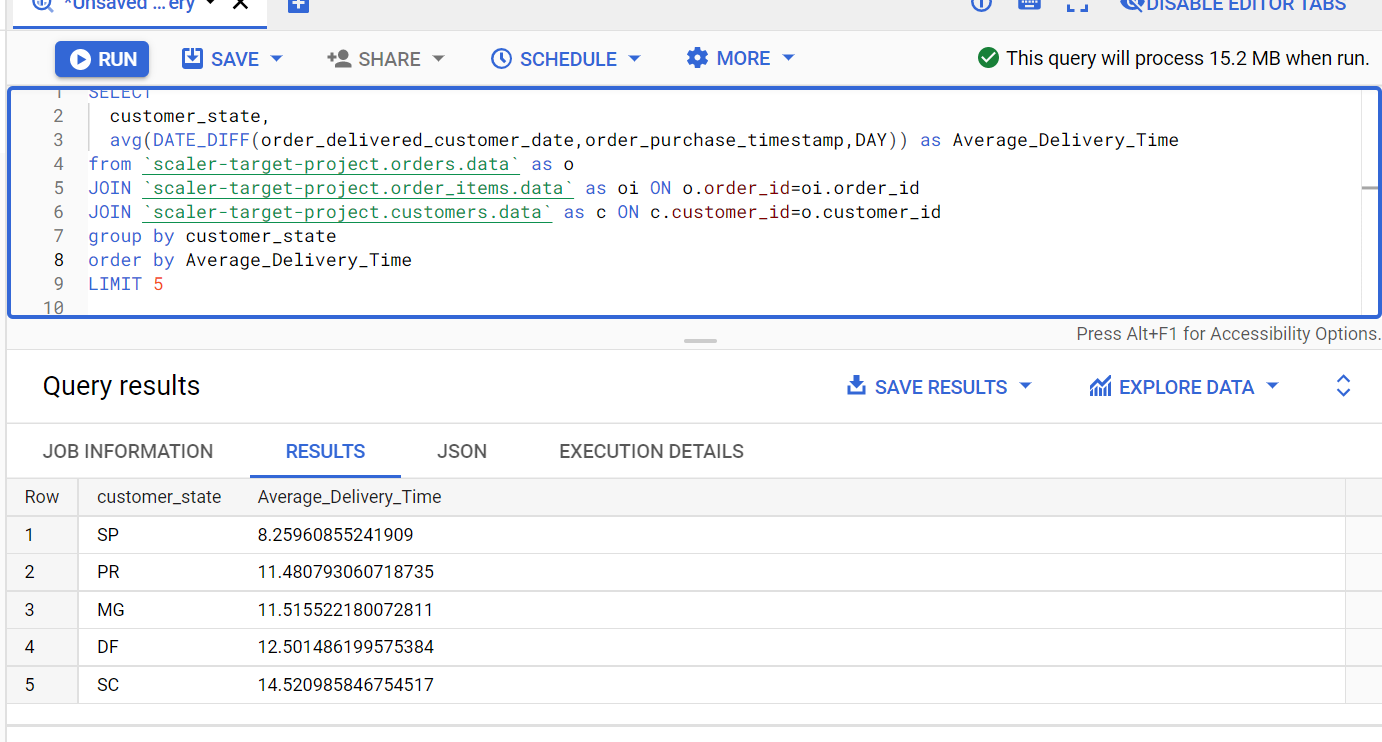
a. Top 5 states with highest/lowest average freight value



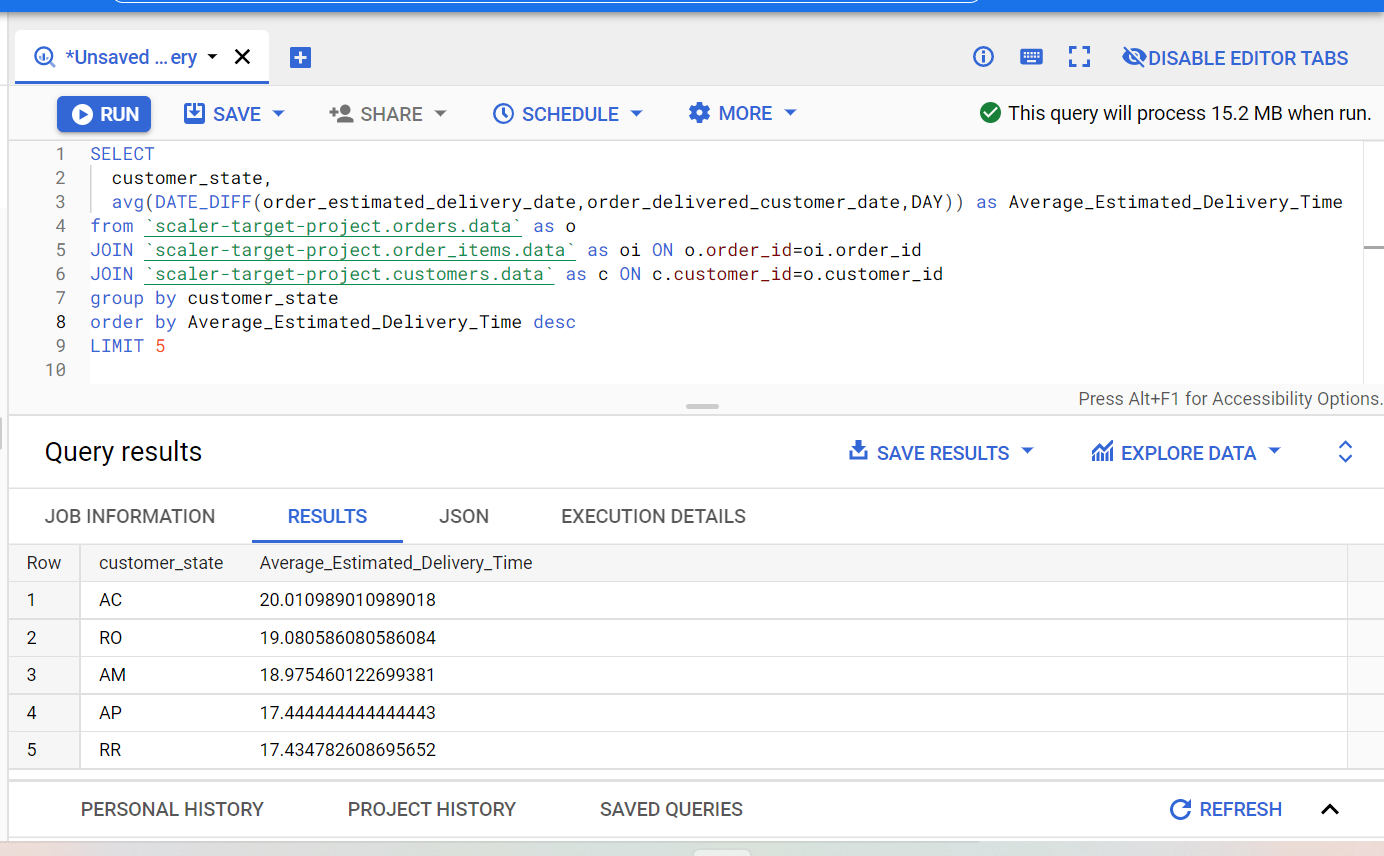


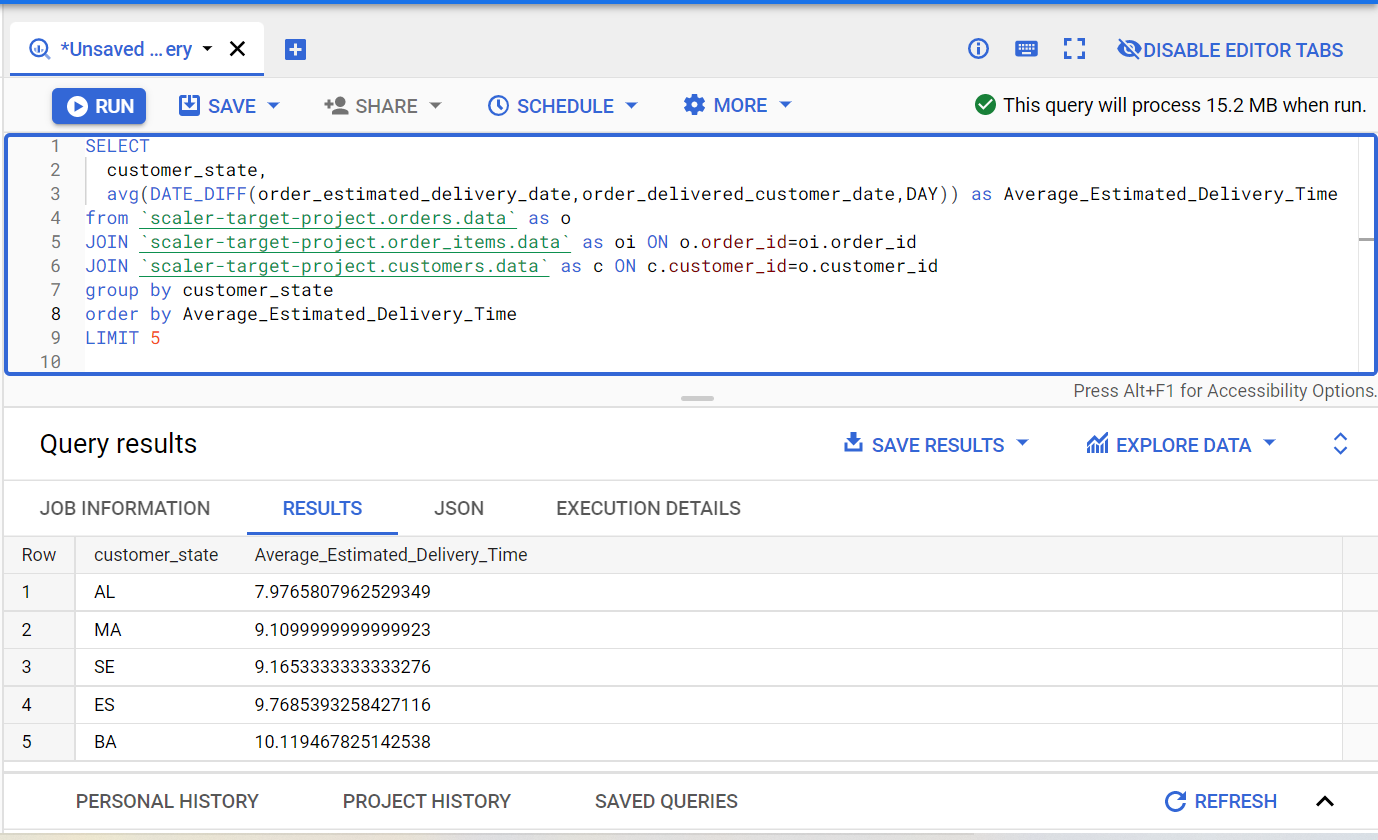
b. Top 5 states with highest/lowest average time to delivery





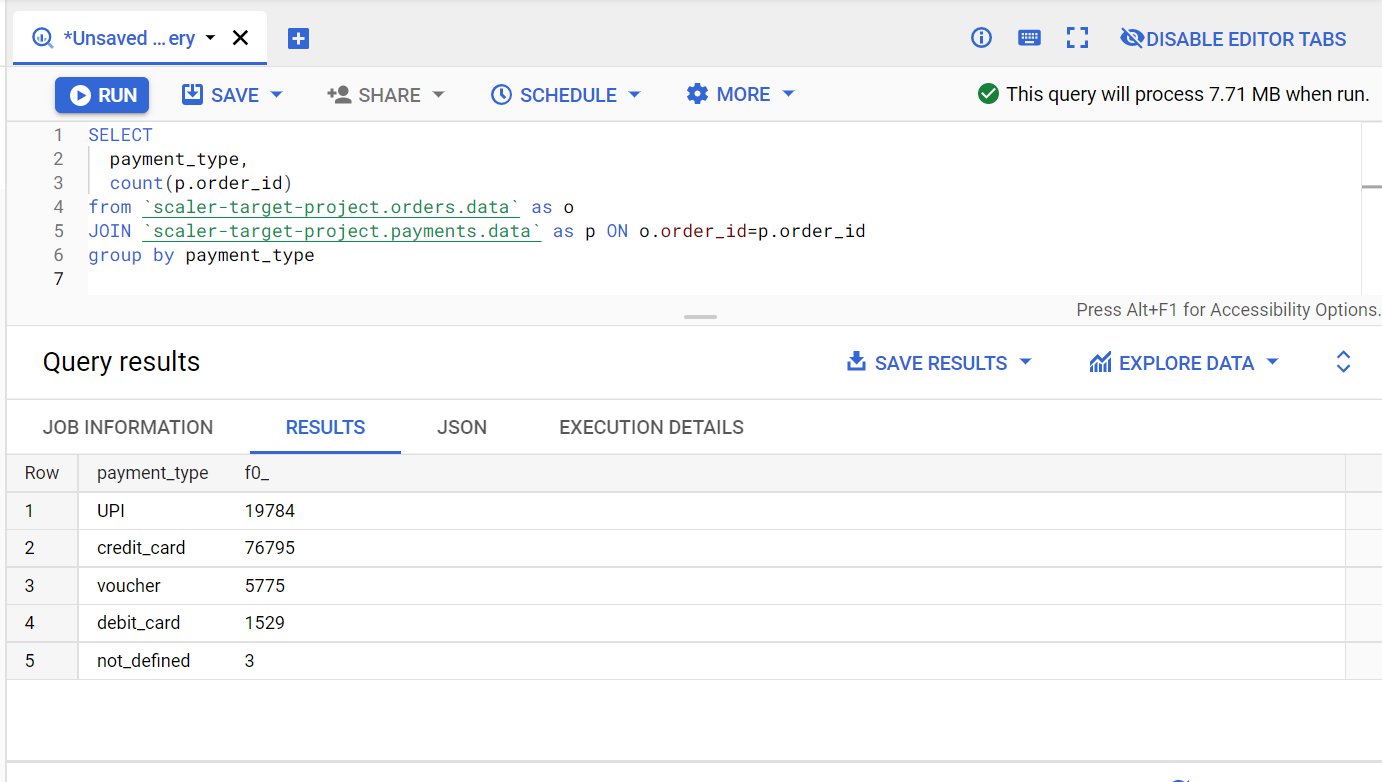
c. Top 5 states where delivery is really fast/ not so fast compared to estimated date





5. Payment type analysis: Join “payments” dataset with the existing data on order\_id

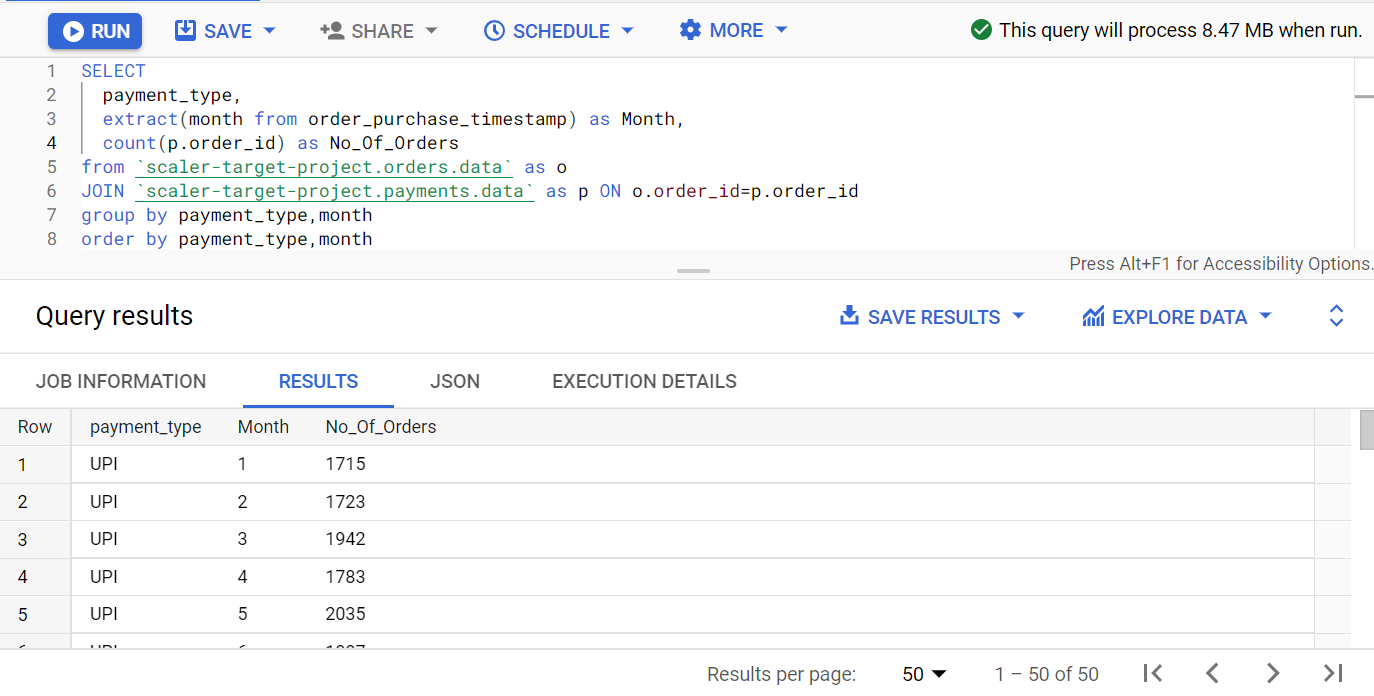
a. Count of orders for different payment types



b. Distribution of payment installments and count of orders



c. Count of orders for different payment types Month over Month



Actionable Insights:

* There are 4119 cities in 27 states to which orders are getting delivered.
* There are 611 cities in 23 states from which sellers sell their goods.
* 609 orders are unavailable.
* The numbers of orders are increasing year in year.
* Brazilians mostly buy on Saturday, Sunday and Monday.
* Brazilians mostly buy in afternoon and night.
* Most orders are placed from states “SP”, “RJ” and “MG”.
* Least orders are placed from “RR”, “AP” and “AC”.
* Most orders are placed from cities “sao Paulo”, “rio de janeiro” and “belo horizonte”
* Highest average Freight value is for states such as “RR”, “PB” and “RO”.
* Lowest average Freight value is for states such as “SP”, “PR” and “MG”.
* Highest average sales is for states such as “AC”, “PB” and “AL”.
* Lowest average sales is for states such as “SP”, “PR” and “RS”.
* Highest average delivery time is for states “RR”, “AP” and “AM”.
* Lowest average delivery time is for states “SP”, “PR” and “MG”.
* Most payments are done using Credit Card and UPI.
* Most people prefer no of installments of 1, 2 and 3.

Recommendations:

* There are total 23 seller states, but no of customer states are 27. More sellers can be identified in the other states so that the freight value can be decreased and also the delivery time.
* With the same logic as above, more sellers can be identified in various other small and big cities to reduce the freight value and also the delivery time as there are 4119 customer states and 611 seller states.
* Target can get more variety of goods and keep abundant stock of all goods since 609 orders were unavailable during the time period.
* Since Brazilians mostly buy on weekends and Monday, therefore Target can provide discounts or offers on weekdays to influence people to buy more on weekdays.
* Since Brazilians mostly buy in afternoon and night, therefore Target can provide discounts or offers in morning to influence people to buy more in mornings.
* More offers and discounts can be offered to people living in states like “RR”, “AP” and “AC” to influence them to buy more since these are the states from where lowest orders are placed.
* More sellers can be identified in state such as “RR”, “PB” and “RO” to decrease the freight value since these are the states with highest average freight value.
* More offers and discounts can be offered to people living in states “SP”, “PR” and “RS” because these are the states where the average sales is the lowest.
* Highest average delivery time is for states “RR”, “AP” and “AM”. More sellers can be identified in these states to bring down the delivery time, which will also help in reducing the freight value.
* Most payments are done using Credit Card. Target can offer some offers to influence more people to pay using UPI.