# SQL Business Case

# 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 allows 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.

**Data Dictionary :**

Data is available in 8 csv files:

1. customers.csv

2. geolocation.csv

3. order\_items.csv

4. payments.csv

5. reviews.csv

6. orders.csv

7. products.csv

8. sellers.csv

Each feature or columns of different CSV files are described below:

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. |

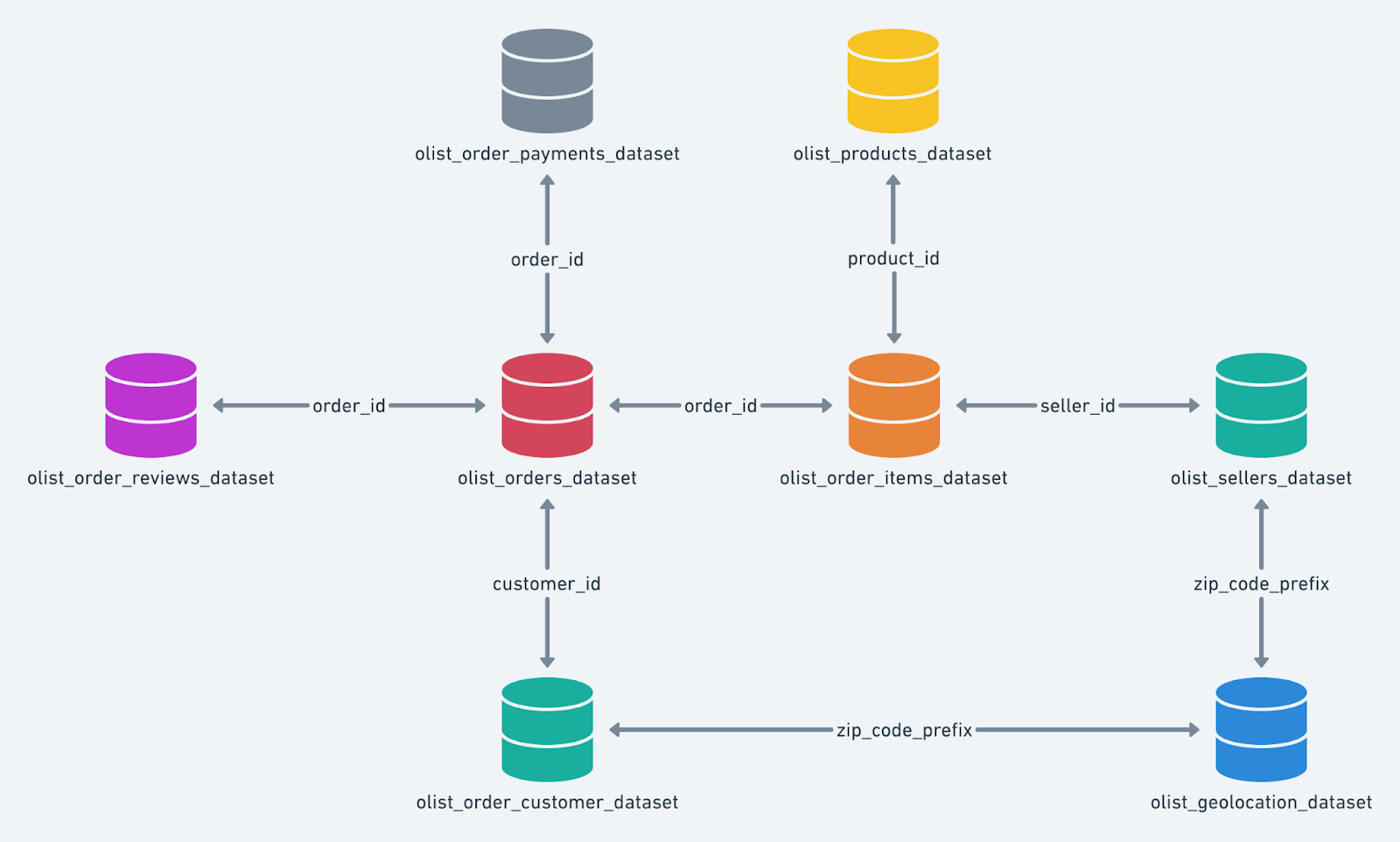
The **reviews.csv** contain following features:

|  |  |
| --- | --- |
| **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. |

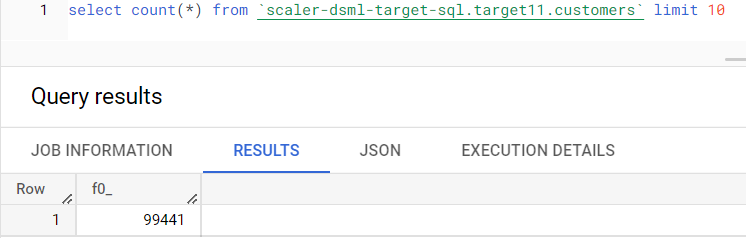
**High level overview of relationship between datasets:**



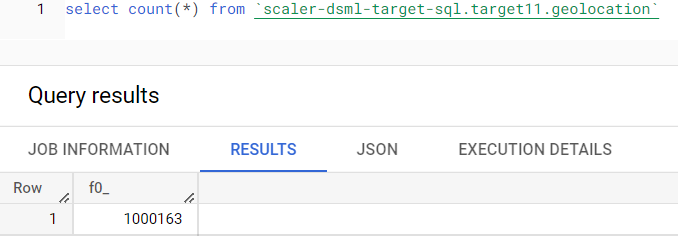
# Import the dataset and do usual exploratory analysis steps like checking the structure & characteristics of the dataset

## Get number of rows in the data

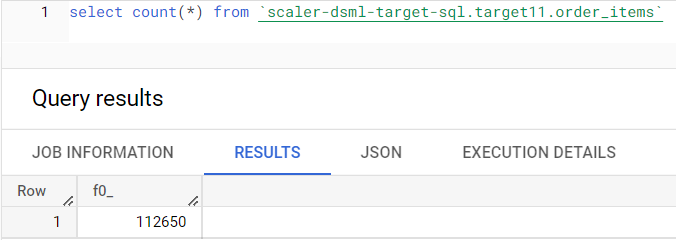
### For Customers Table :



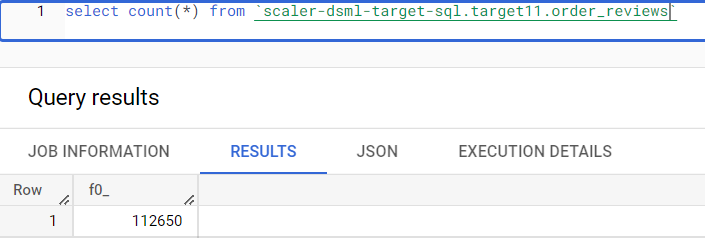
### For GeoLocation Table:



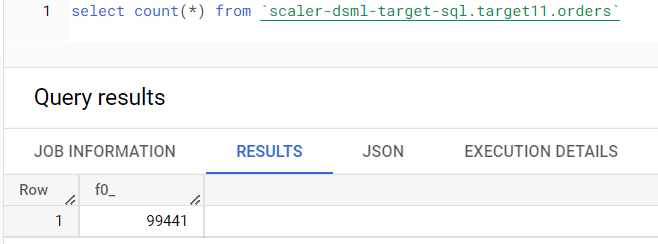
### For order\_items table:



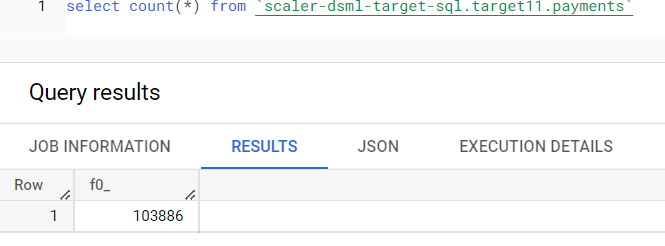
### For order\_reviews table:



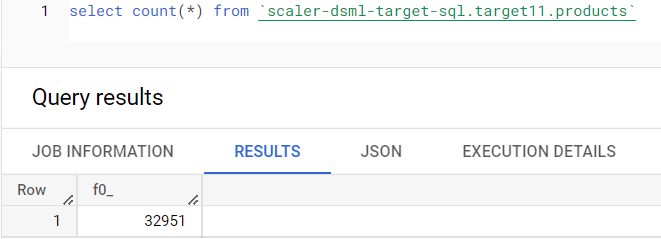
### For orders table :



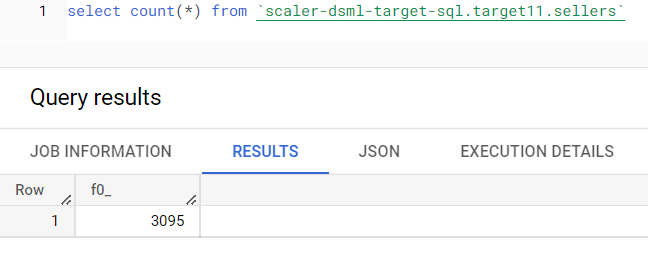
### For payments table:



### For products table :

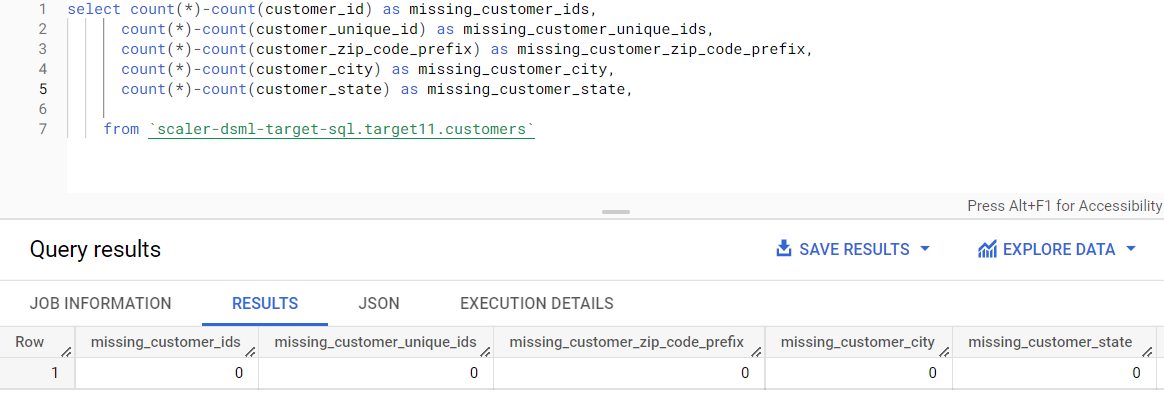


### For sellers table :

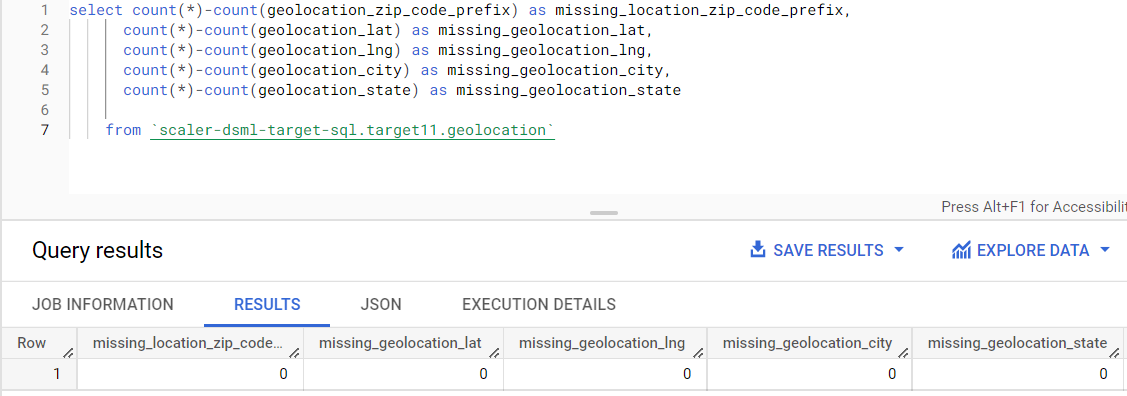


## Number of null or missing values in a column

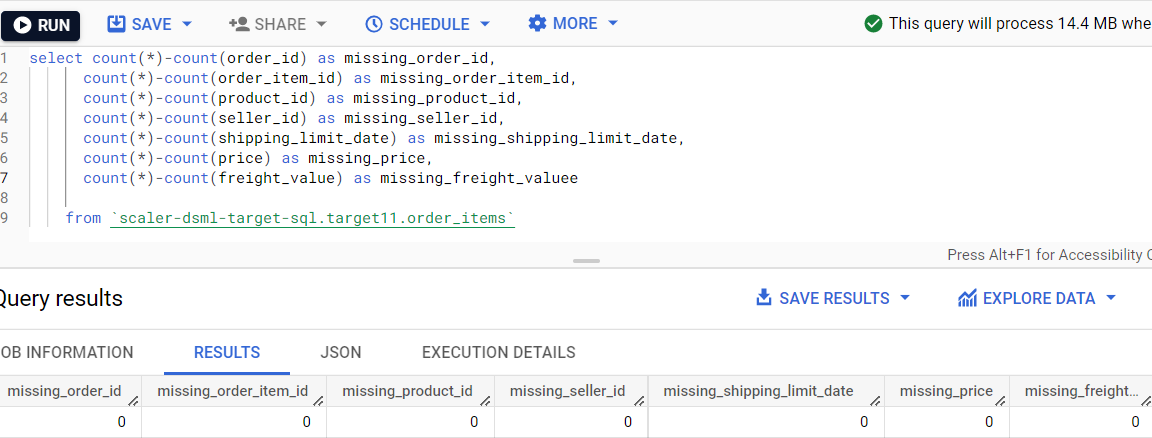
### For Customers Table :



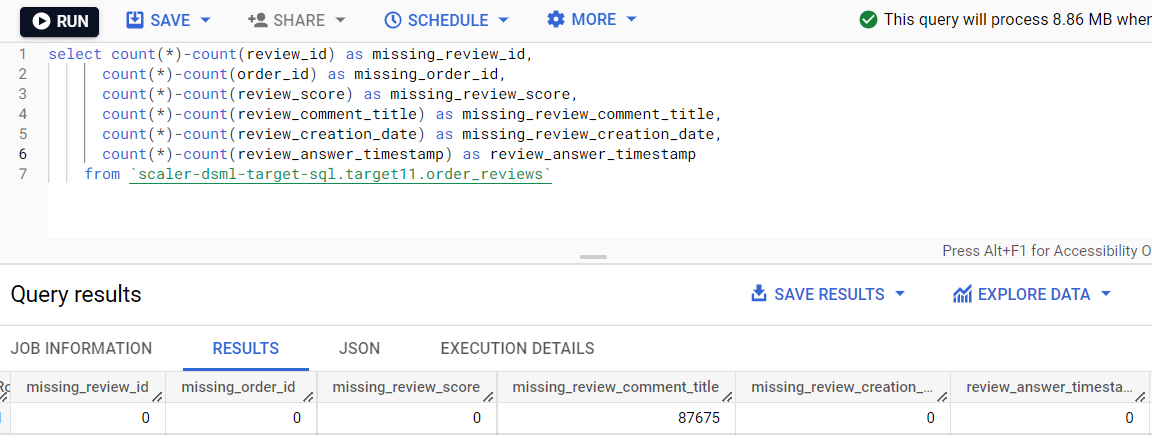
### For GeoLocation Table:



### For Order\_items Table:

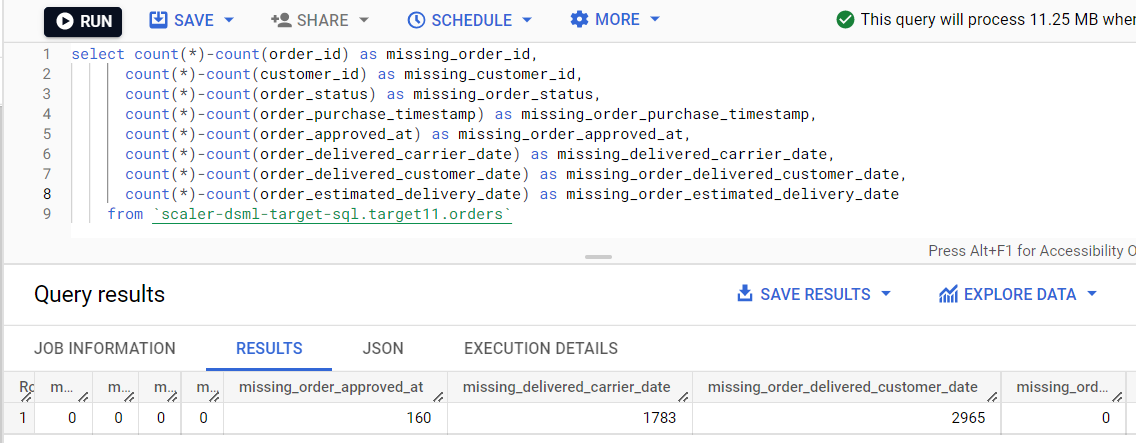


### For Order\_reviews Table :



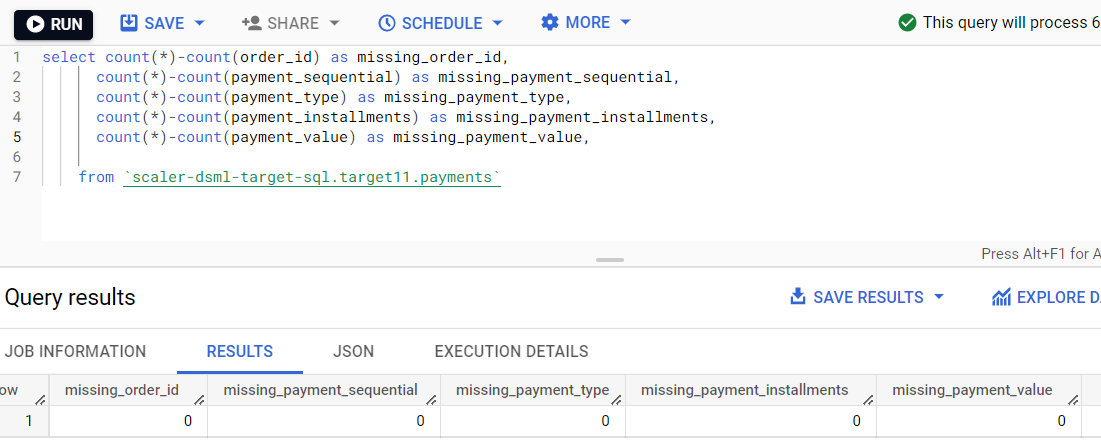
**There are 87675 missing values of Review\_comment\_title**

### For Orders table :

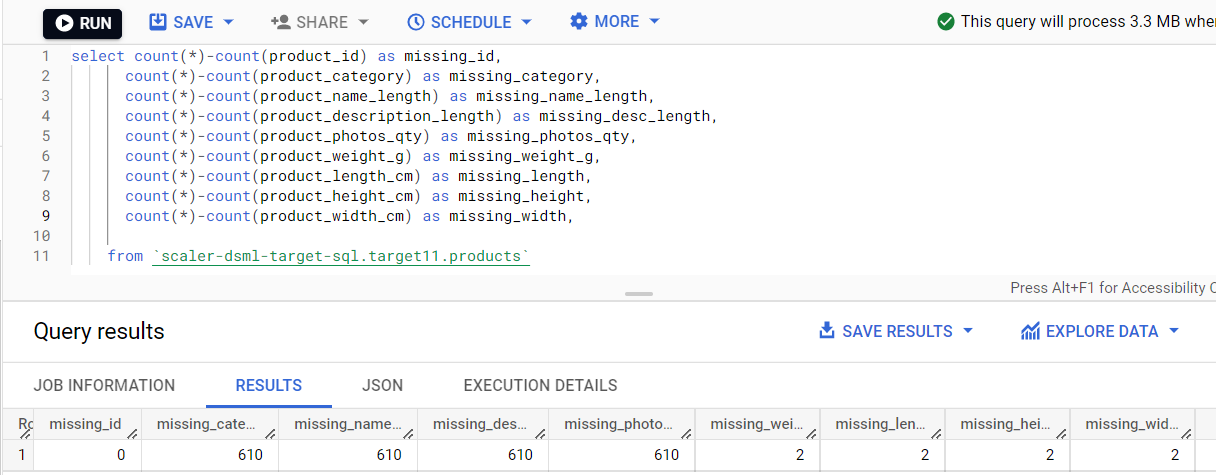


**There are 160 missing values in order\_approved\_at column, 1783 missing values in order\_delivered\_carrier\_date and 2965 missing values in order\_estimated\_delivery\_date**

### For Payments table :

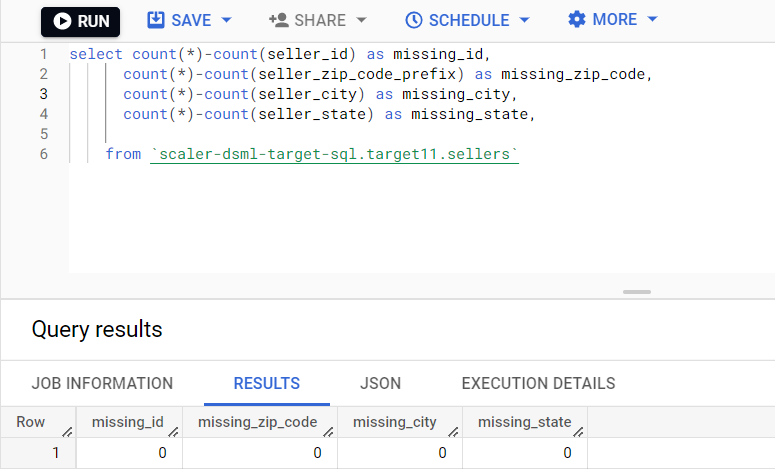


### For Products table :

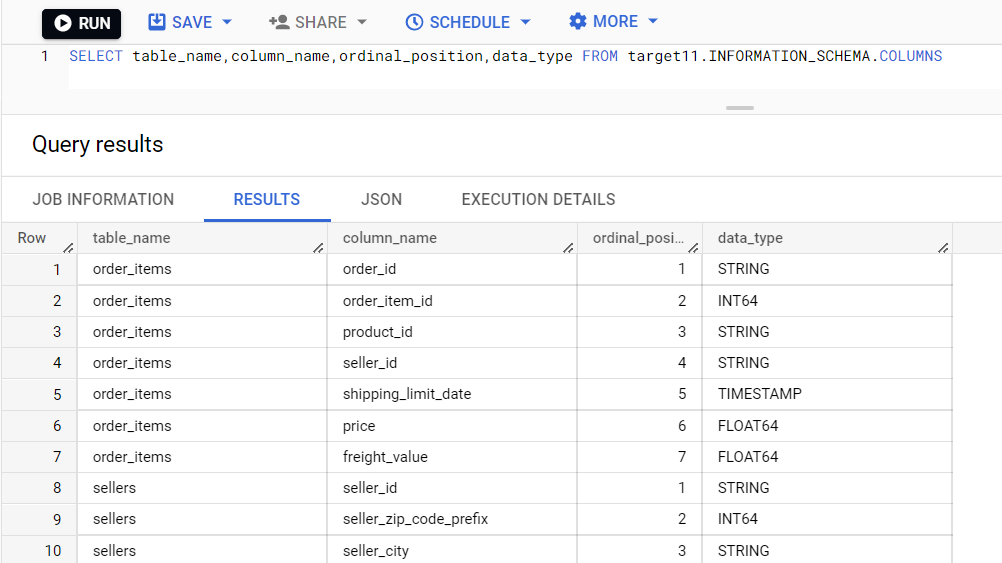


**There are 610 missing values in each columns of product\_category, product\_name\_length, product\_description\_length, product\_photos\_qty and 2 missing values in each of product\_length\_cm, product\_height\_cm and product\_width\_cm.**

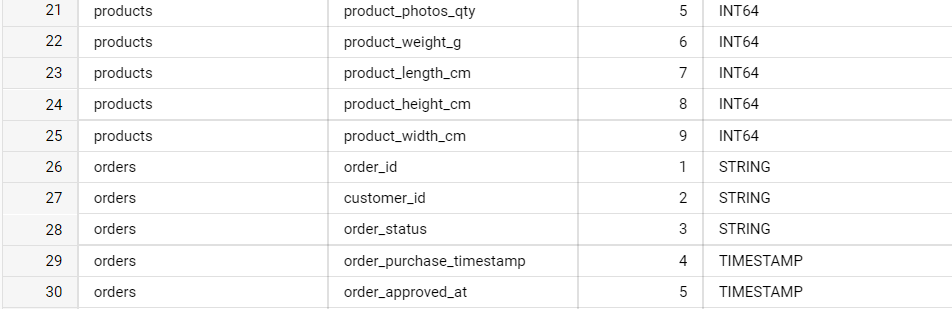
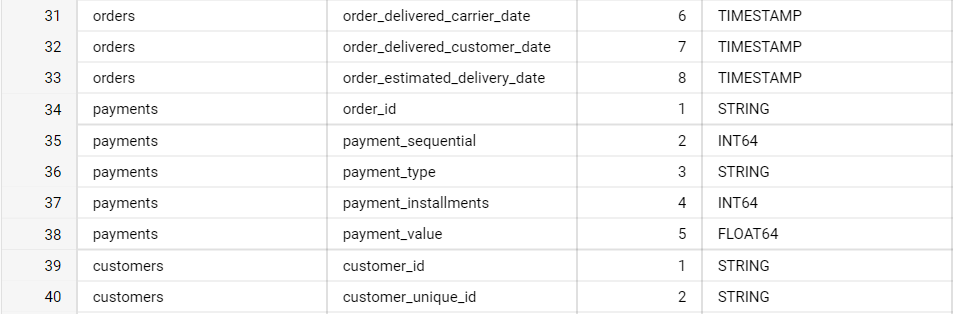
### For Sellers table :



## Data type of columns in a table

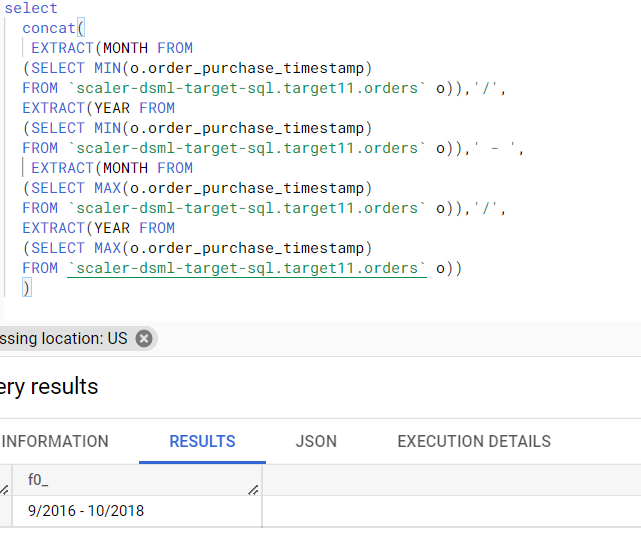




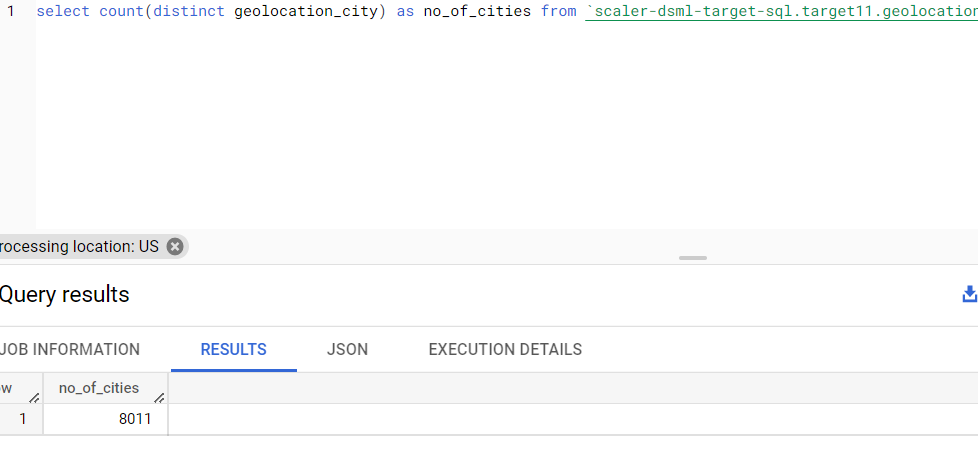
 



## Get the time period for which the data is given



## Number of cities in our dataset

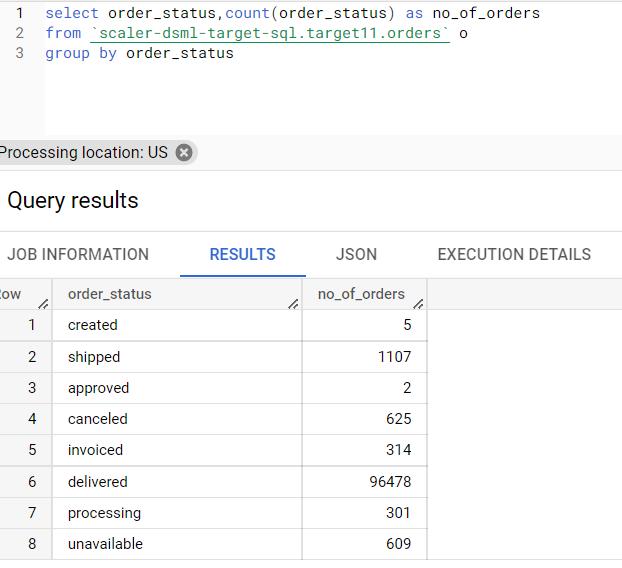


## Number of states in our dataset

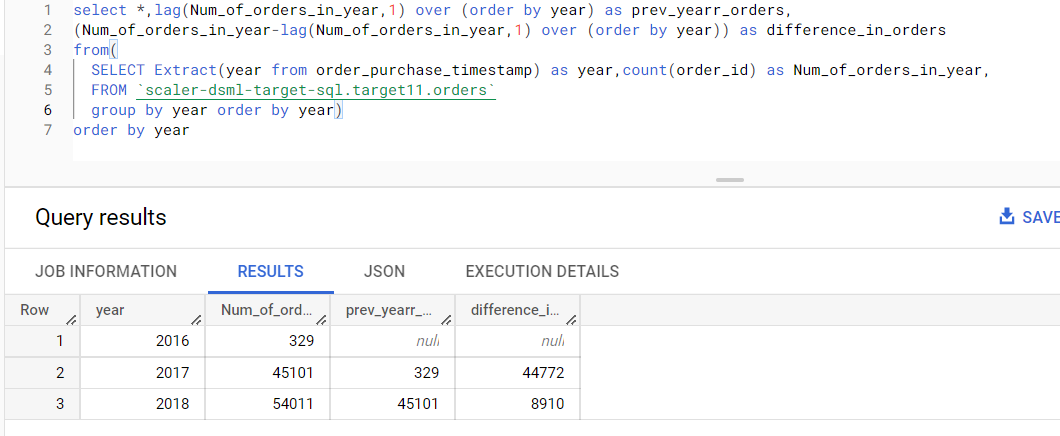


# In-depth Exploration:

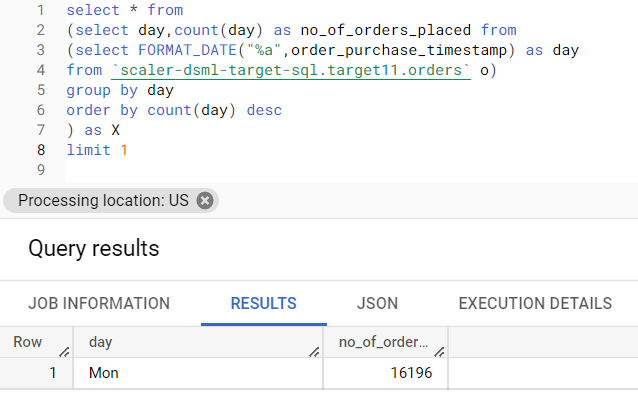
## How many orders do we have for each order status?



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



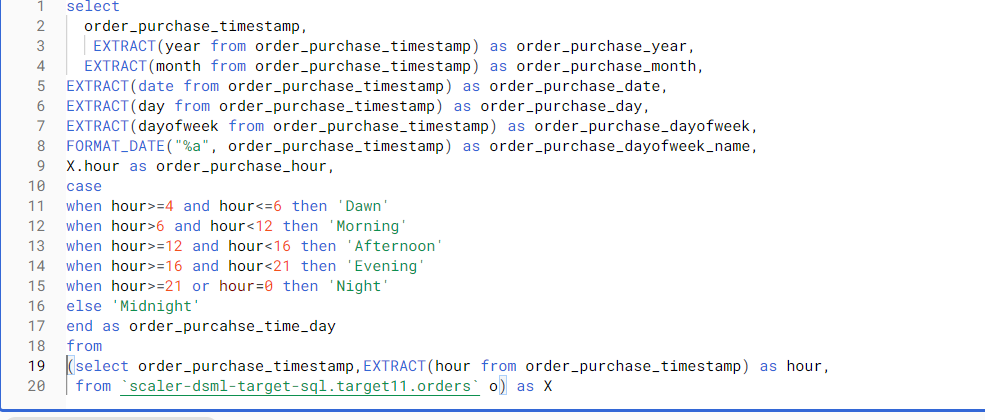
## On what day of week brazilians customers tend to do online purchasing?

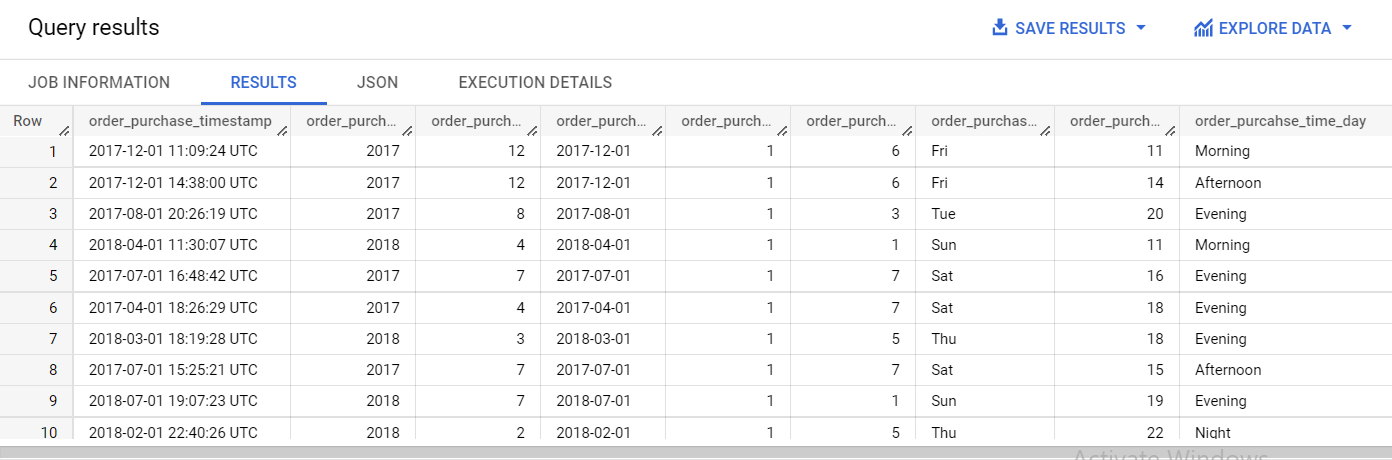


## What time do Brazilian customers tend to buy (Dawn, Morning, Afternoon or Night)?



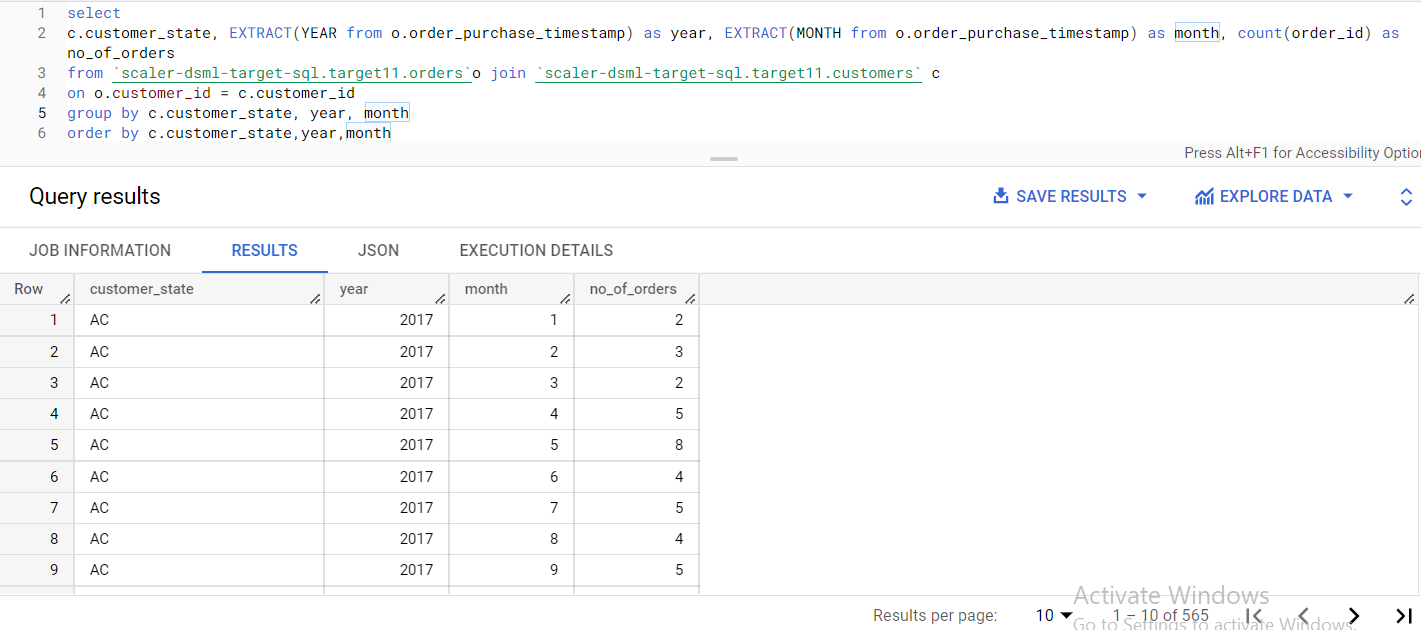
1. Feature Extraction: Through order\_purchase\_timestamp in “orders” dataset extract
   * 1. order\_purchase\_year
     2. order\_purchase\_month
     3. order\_purchase\_date
     4. order\_purchase\_day
     5. order\_purchase\_dayofweek
     6. order\_purchase\_dayofweek\_name
     7. order\_purchase\_hour
     8. order\_purchase\_time\_day



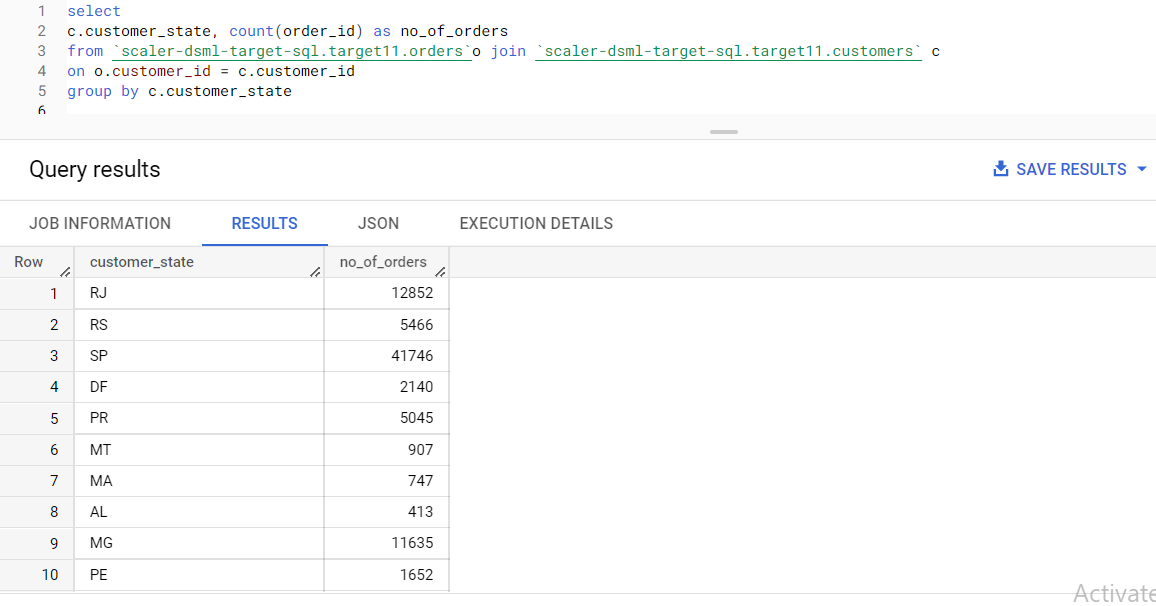


# Evolution of E-commerce orders in the Brazil region:

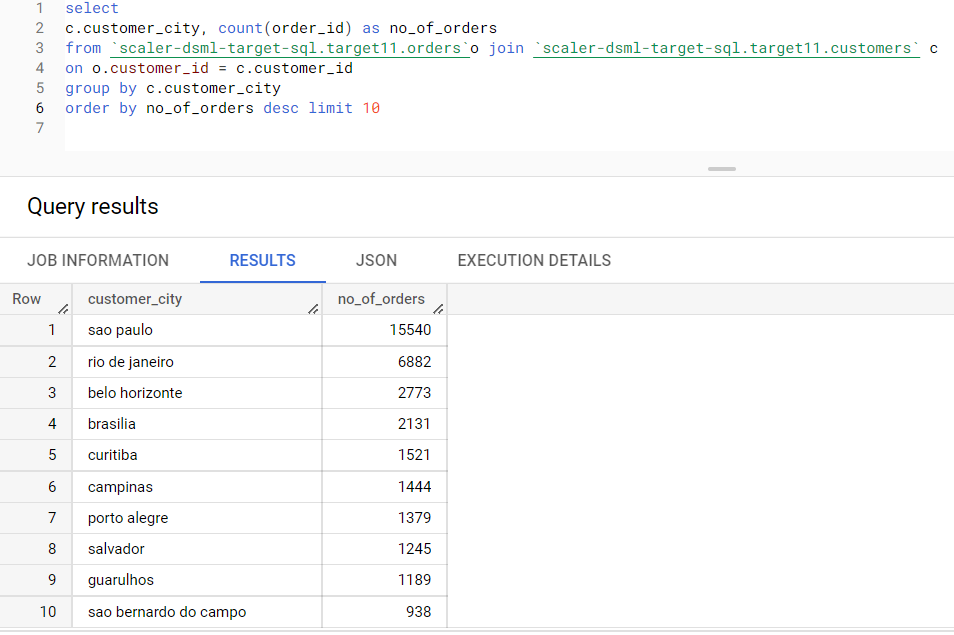
### Get month on month orders by region



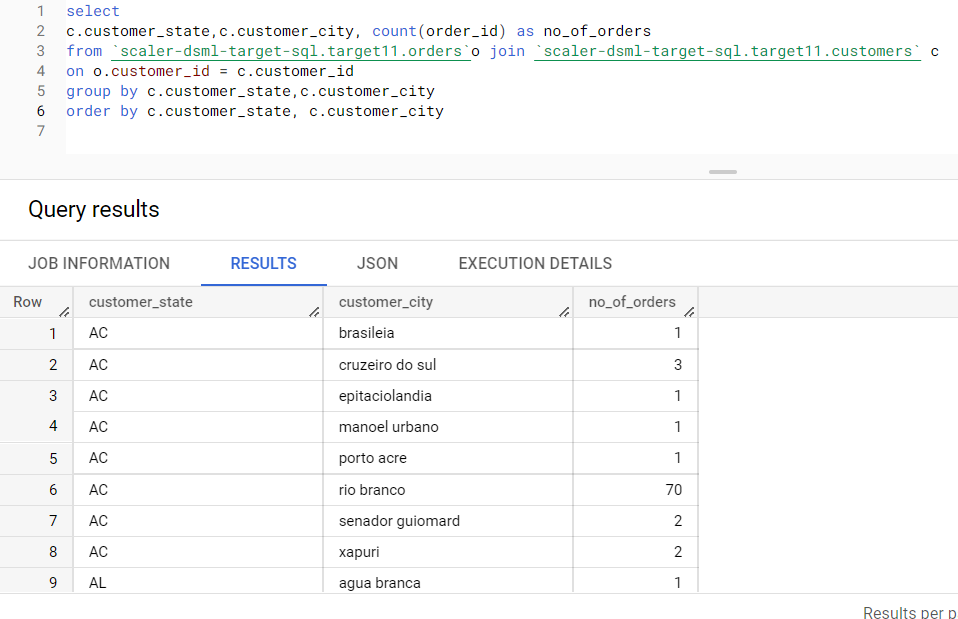
## Total of customer orders by state



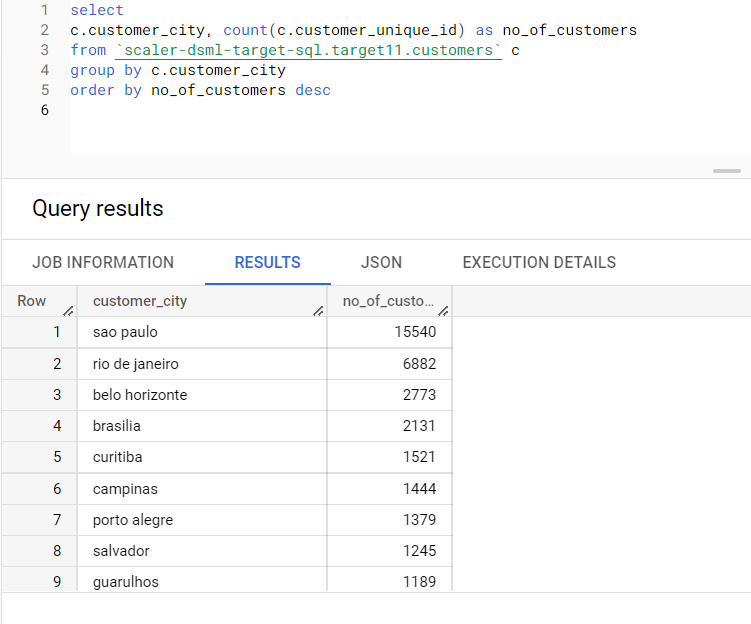
## Top 10 brazilian cities most no. of orders



## How are customers distributed in Brazil



## 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)

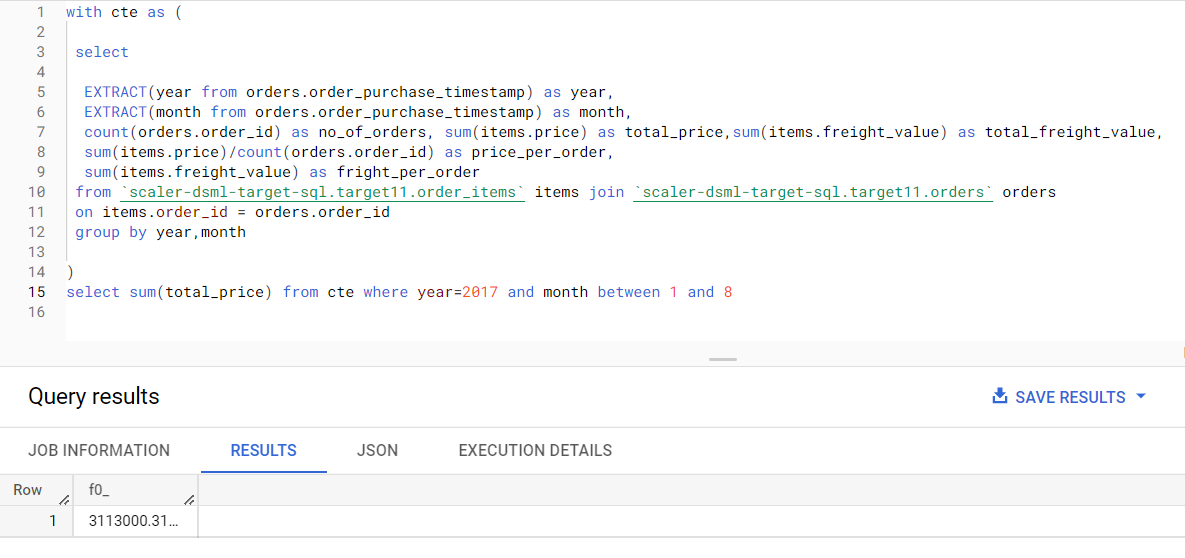
3. Create new columns:

price\_per\_order = sum(price) / count(order\_id)

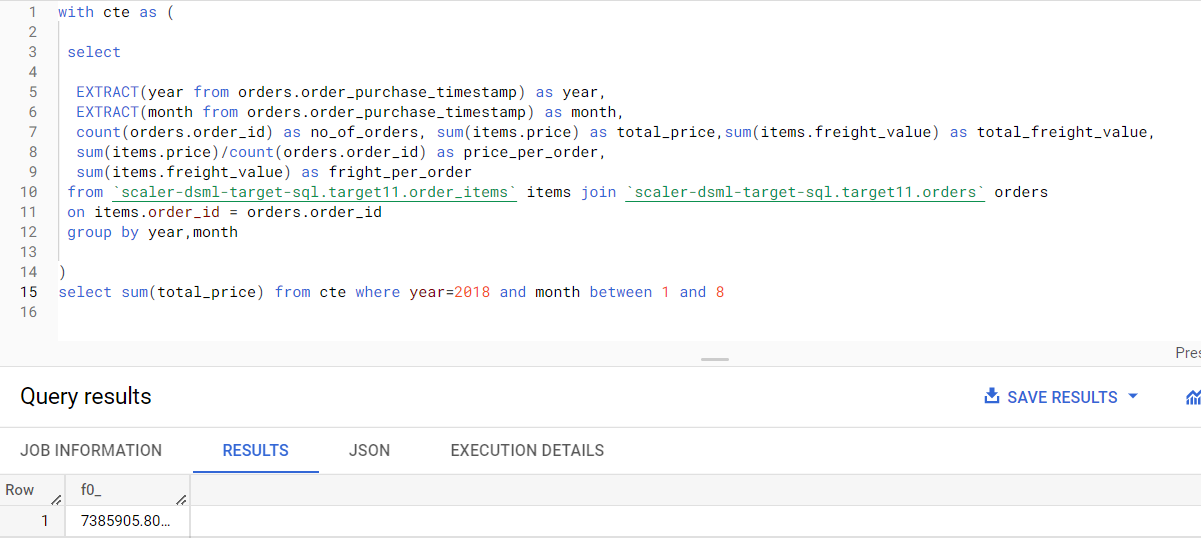
freight\_per\_order= sum(freight\_value) / count(order\_id)

## Step 2: Answer the following questions:

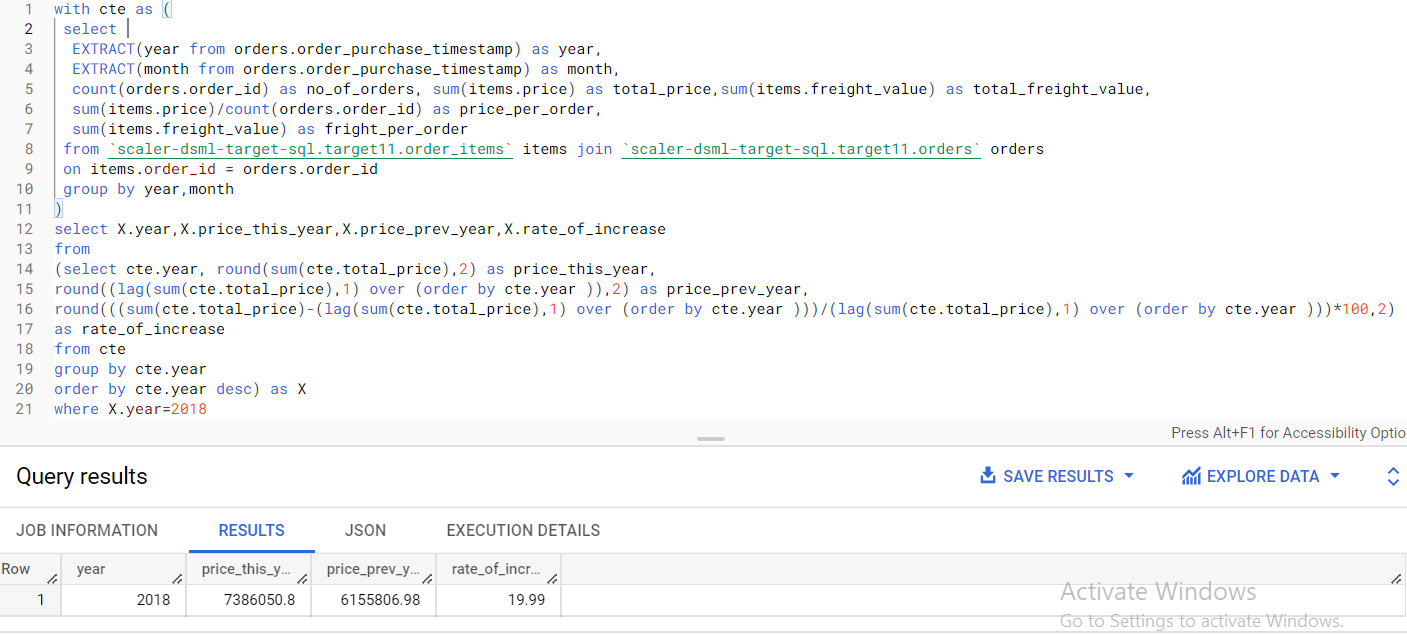
### Total amount sold in 2017 between Jan to August



### Total amount sold in 2018 between Jan to august

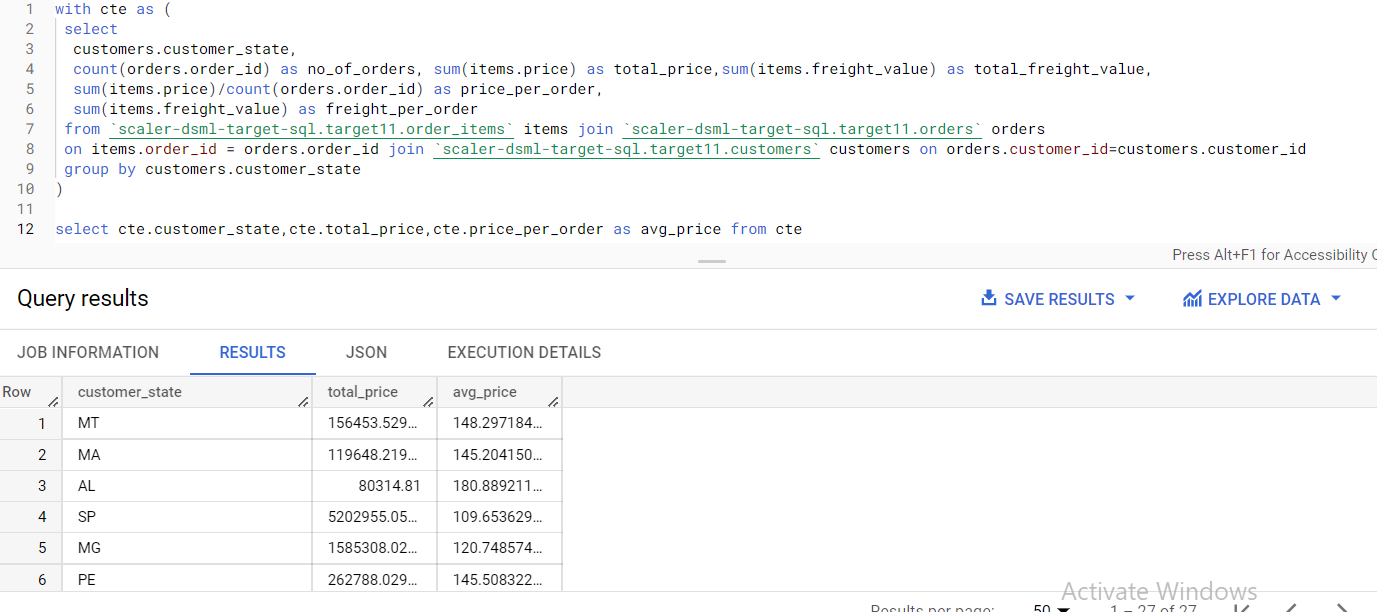


### % increase from 2017 to 2018

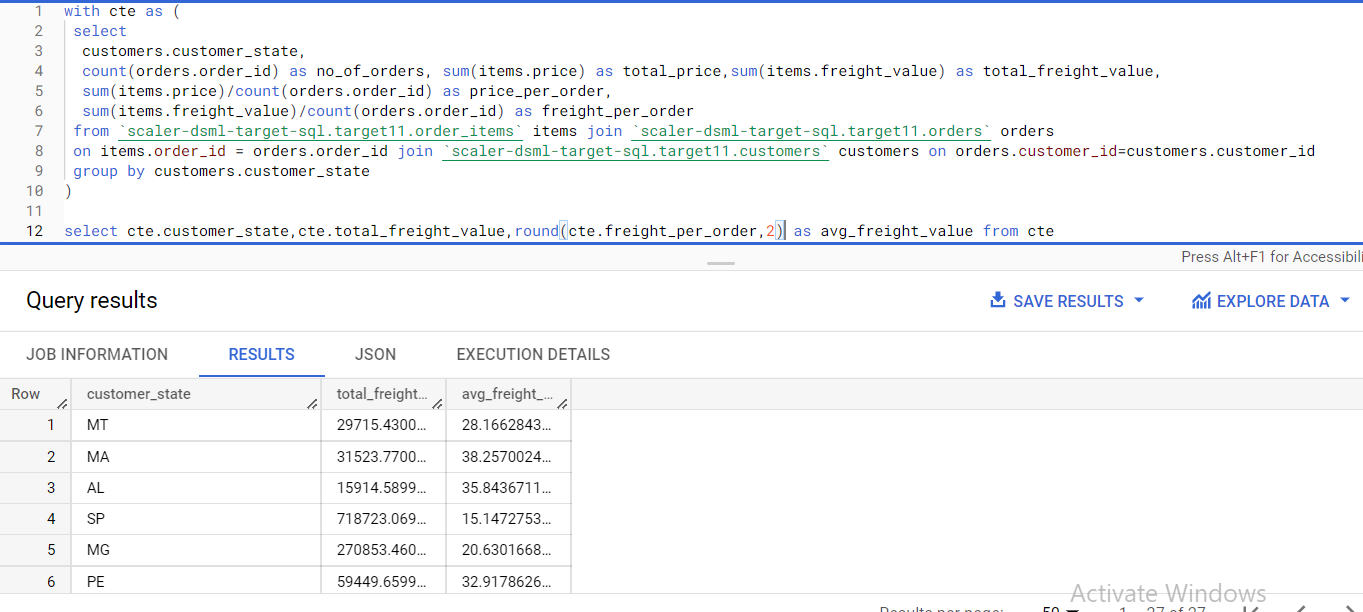


## Step 3: Join (orders+order\_items) table from previous step with “customers” table on Customer\_id and find:

### Mean & Sum of price by customer state



### 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

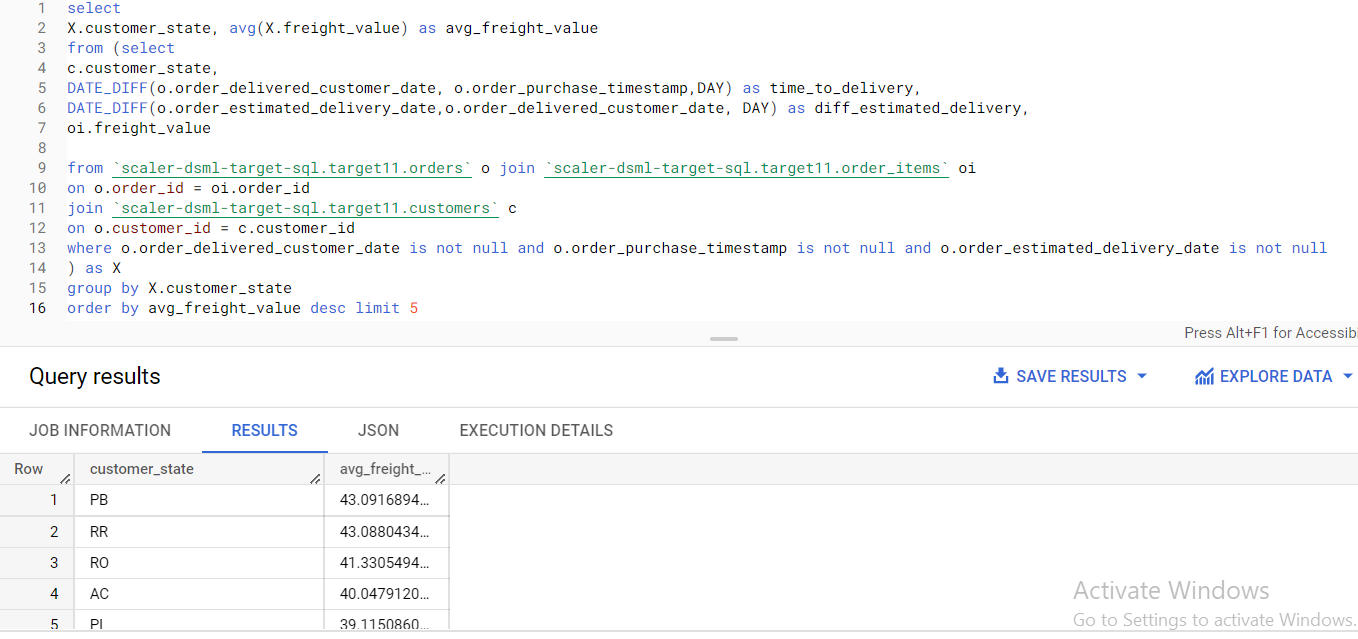
diff\_estimated\_delivery = order\_estimated\_delivery\_date-order\_delivered\_customer\_date

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

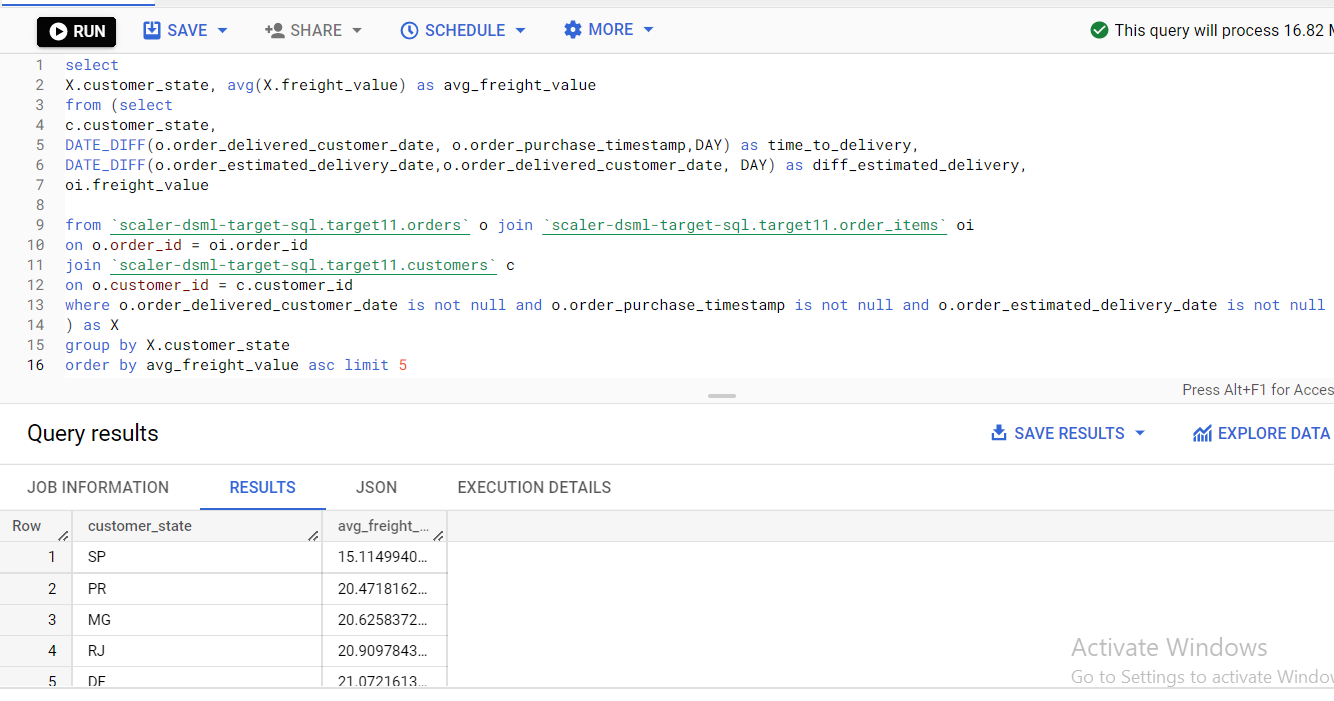
4. Sort the data to get the following:

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

States with Highest freight value :

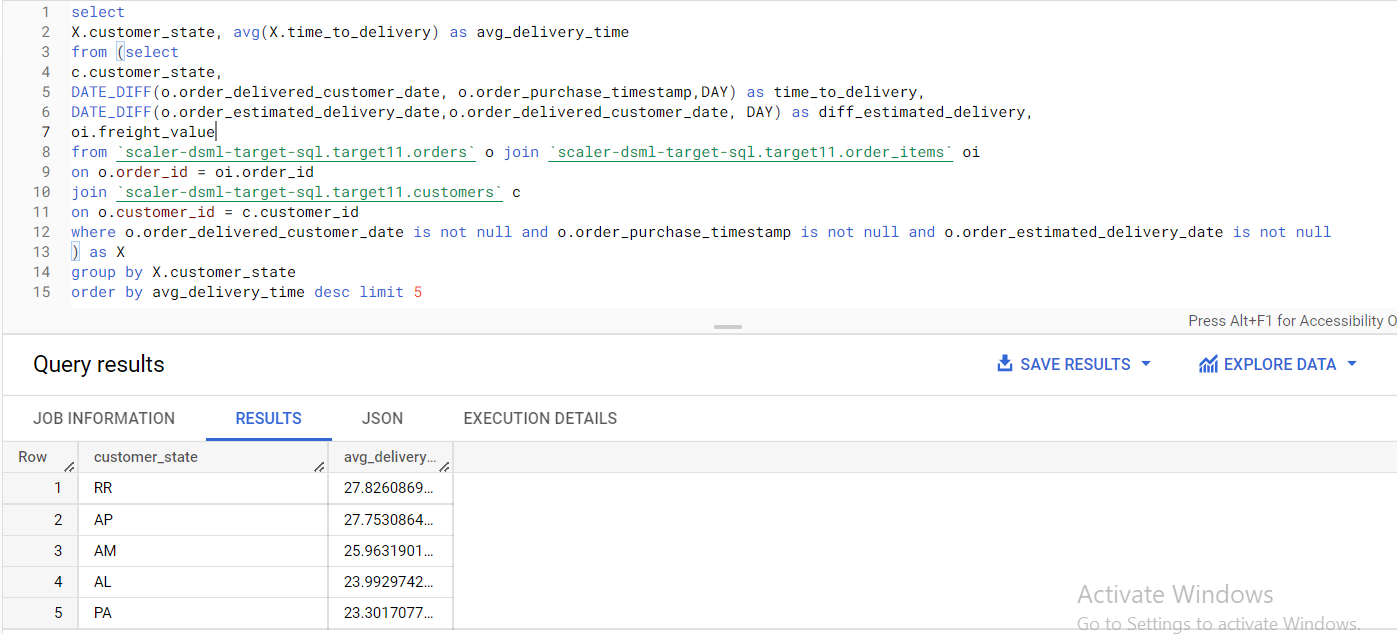


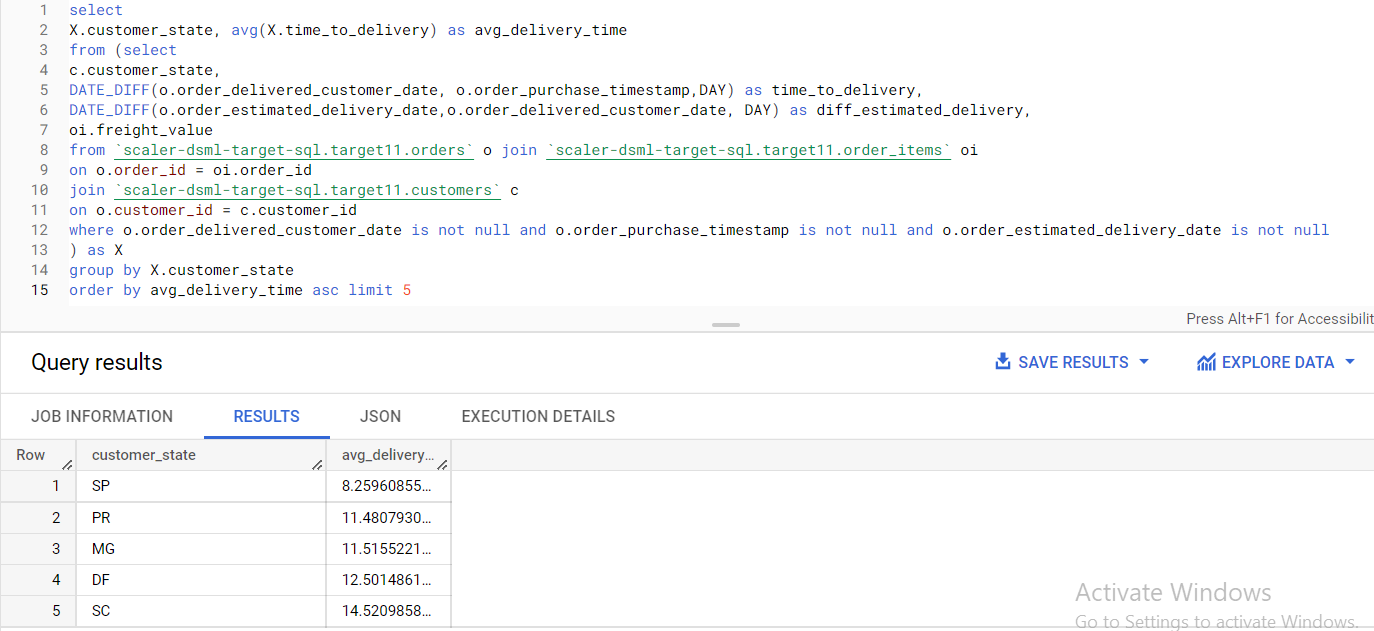
**States with Lowest freight Value :**



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

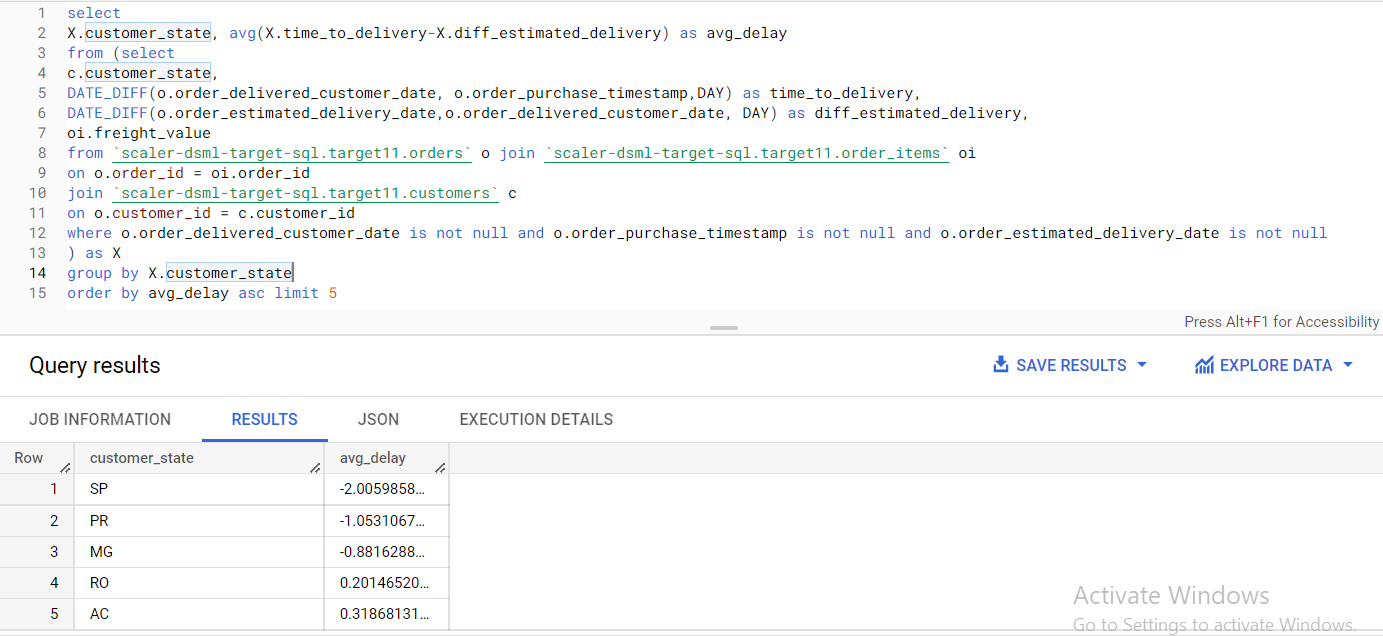
**States with highest avg delivery time :**





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

**states with delivery fast compared to estimated delivery :**



**states with late delivery compared to estimated delivery :**

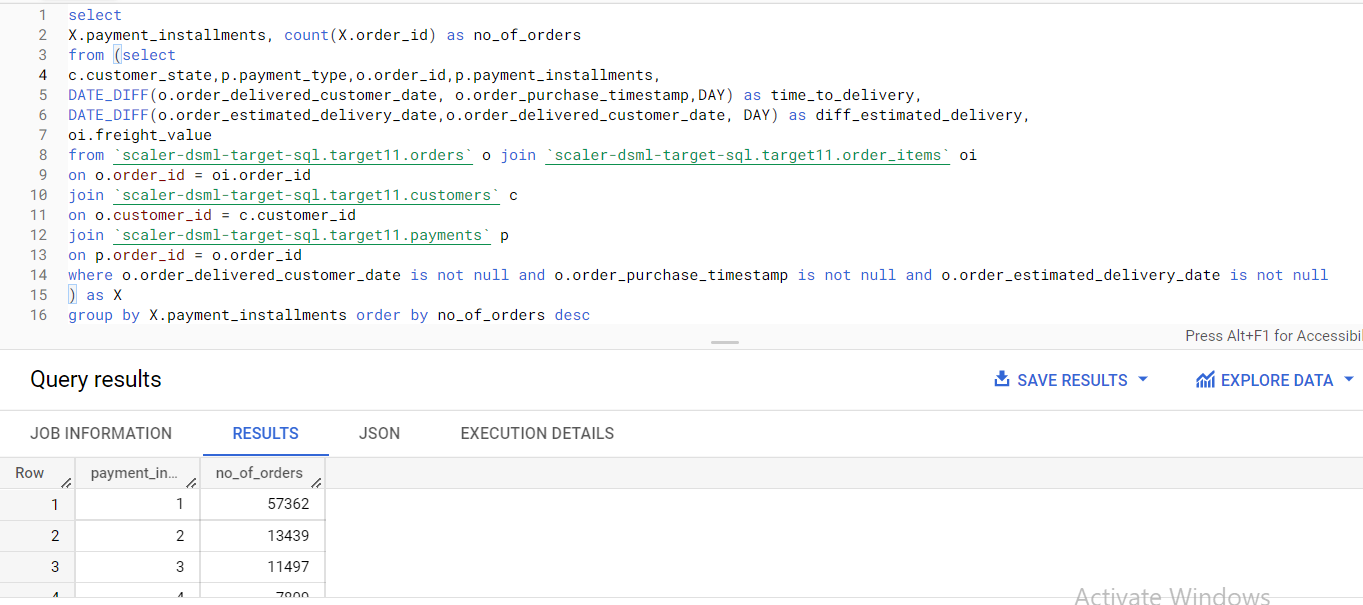


# 6. Payment type analysis: Join “payments” dataset with the existing data on order\_id

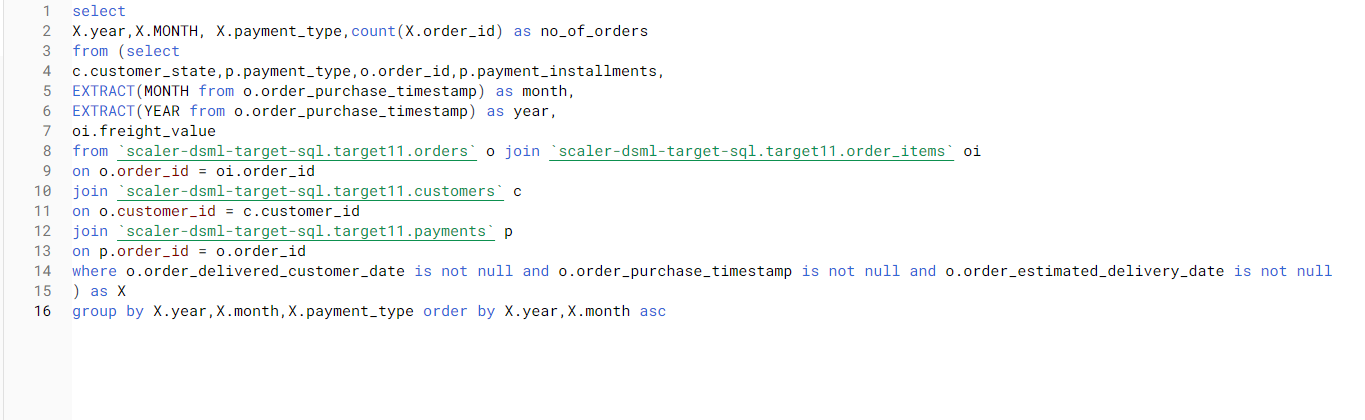
a. Count of orders for different payment types

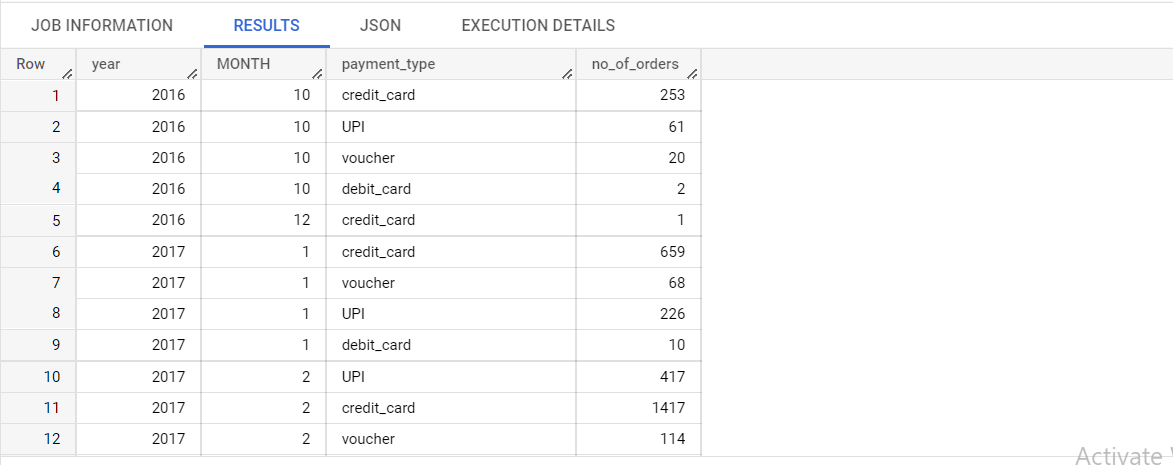


b. Distribution of payment installments and count of orders

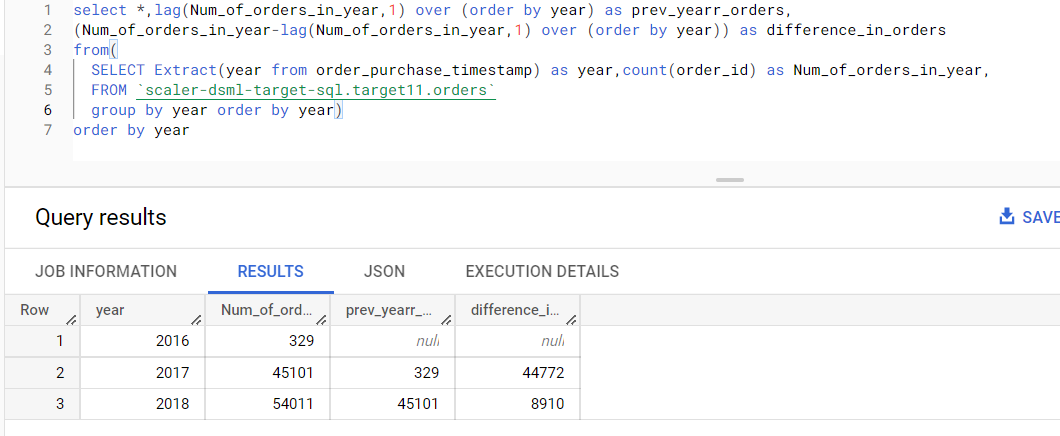
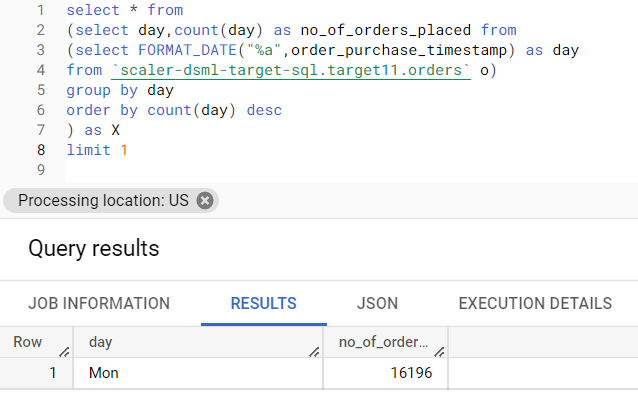


c. Count of orders for different payment types Month over Month





# Insights :

* There are 99441 orders placed in the given data
* There are 32951 different types of products which the customers ordered
* There are 87675 reviews where title is missing for review\_comments.
* There are 160 missing values in order\_approved\_at, 1783 missing values in order\_delivered\_delivered\_carrier\_date and 2965 missing values in order\_estimated\_delivery\_date in Orders table
* There are 610 missing values in each columns of product \_category, product name product description\_length, product photos qty and 2 missing values in each of product length\_ cm, product\_height \_cm and product width\_cm.
* The data is between September 2016 and October 2018.
* There are 27 states and 8011 cities in the data
* There is growing trend in ecommerce usage over the years. 
* More number of orders are placed on Monday
* The states Sp, RJ and Mg have more number of orders while the states AC,AP and AR have least number of orders
* The “sau palo” city has more number of orders
* There is 19.99% of revenue increase from 2017 to 2018 even when the data of 2018 is till October only
* The states PB and SP have highest and lowest avg freight rates respectively.
* The states RR and SP have highest and lowest avg delivery times.
* SP state has the lowest delay where as AL state has highest order delay between estimated and delivered time
* More number of orders are placed through credit card
* More number of orders are paid in one installment 

# Recommendations :

* Every time an order is placed by a customer, each instance is treated as different customer but there will be a high possibility that there are repeated customers. If each instance of order is treated as different customer, it is difficult to analyze the customer behavior.
* There are 2965 missing values in estimated\_delivery\_time which is not a good experience for a customer to trace or expect the order
* The data is available for only four months i.e.(September to December) of data in 2016 and only ten months i.e.(January to October) in 2018. It’s better to have data for whole year so that we can have proper estimation of trends over years.
* The states AC, AP and AR has least number of orders. We can attract more customers by having more campaigns and offers specific to those states.
* The RR state has more avg time taken for delivery. We can decrease the time by having more physical stores and delivery partners.
* The AL state has more difference in delay between the expected delivery and actual delivery time. Need to trace out the reasons for that.
* More number of orders are placed in single installments. We can have more number of orders if we offer no cost EMI