**Table of Contents**

[Import the dataset and do usual exploratory analysis steps like checking the structure & characteristics of the dataset 1](#__RefHeading___Toc703_2854059815)

[Data type of all columns in the "customers" table. 2](#__RefHeading___Toc705_2854059815)

[Get the time range between which the orders were placed 4](#__RefHeading___Toc707_2854059815)

[Count the Cities & States of customers who ordered during the given period 4](#__RefHeading___Toc709_2854059815)

[In-depth Exploration 6](#__RefHeading___Toc711_2854059815)

[Is there a growing trend in the no. of orders placed over the past years 6](#__RefHeading___Toc713_2854059815)

[Can we see some kind of monthly seasonality in terms of the no. of orders being placed? 6](#__RefHeading___Toc715_2854059815)

[During what time of the day, do the Brazilian customers mostly place their orders 7](#__RefHeading___Toc717_2854059815)

[Evolution of E-commerce orders in the Brazil region 8](#__RefHeading___Toc719_2854059815)

[Get the month on month no. of orders placed in each state 8](#__RefHeading___Toc721_2854059815)

[How are the customers distributed across all the states 9](#__RefHeading___Toc723_2854059815)

[Impact on Economy: Analyze the money movement by e-commerce by looking at order prices, freight and others. 10](#__RefHeading___Toc725_2854059815)

[Get the % increase in the cost of orders from year 2017 to 2018 (include months between Jan to Aug only) 10](#__RefHeading___Toc727_2854059815)

[Calculate the Total & Average value of order price for each state. 11](#__RefHeading___Toc729_2854059815)

[Calculate the Total & Average value of order freight for each state. 12](#__RefHeading___Toc731_2854059815)

[Analysis based on sales, freight and delivery time 13](#__RefHeading___Toc733_2854059815)

[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. 13](#__RefHeading___Toc735_2854059815)

[Find out the top 5 states with the highest & lowest average freight value 14](#__RefHeading___Toc737_2854059815)

[Find out the top 5 states with the highest & lowest average delivery time. 15](#__RefHeading___Toc739_2854059815)

[Find out the top 5 states where the order delivery is really fast as compared to the estimated date of delivery. 16](#__RefHeading___Toc741_2854059815)

[Analysis based on the payments 17](#__RefHeading___Toc743_2854059815)

[Find the month on month no. of orders placed using different payment types. 17](#__RefHeading___Toc745_2854059815)

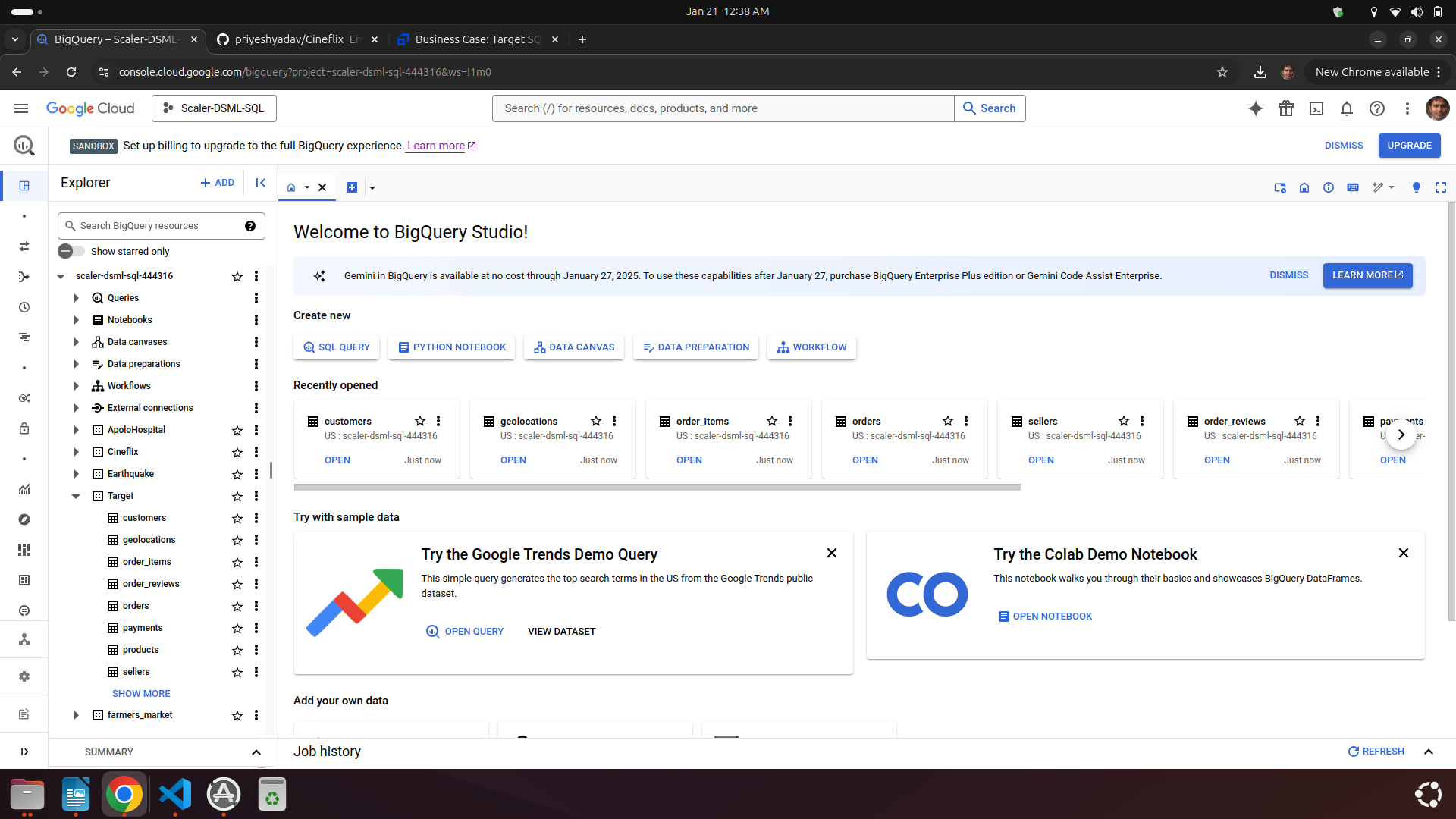
[Find the no. of orders placed on the basis of the payment installments that have been paid 18](#__RefHeading___Toc747_2854059815)

[Actionable Recommendation 19](#__RefHeading___Toc749_2854059815)

TARGET SQL – BUSINESS CASE SOLUTION

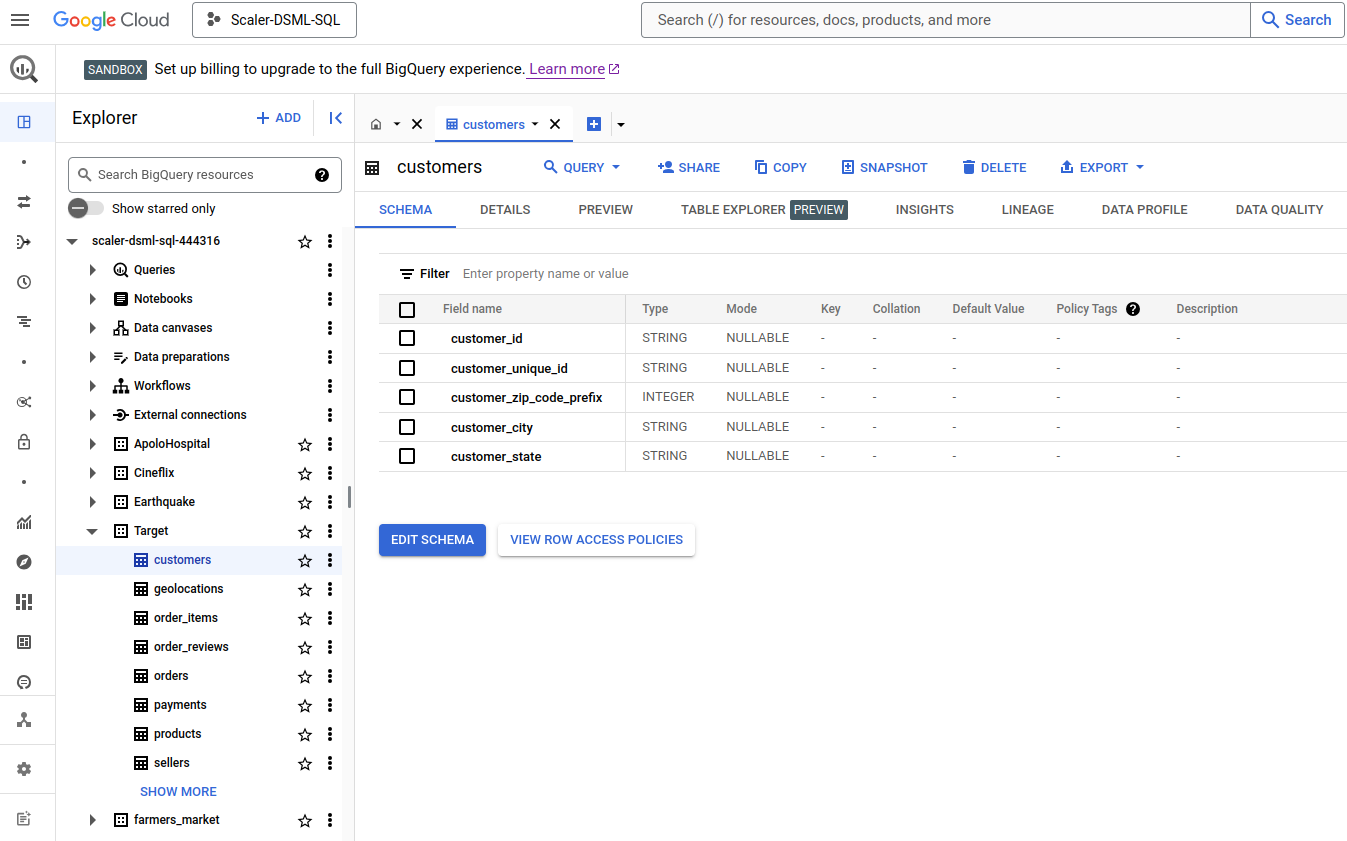
## Import the dataset and do usual exploratory analysis steps like checking the structure & characteristics of the dataset

* Download all(8) CSV files from Google Drive.
* Create a *New Dataset* called ***Target*** in Google BigQuery.
* Create tables under the Target dataset by uploading each CSV one by one, using ***Upload*** & ***Auto Detect*** schema options.



### Data type of all columns in the "customers" table.

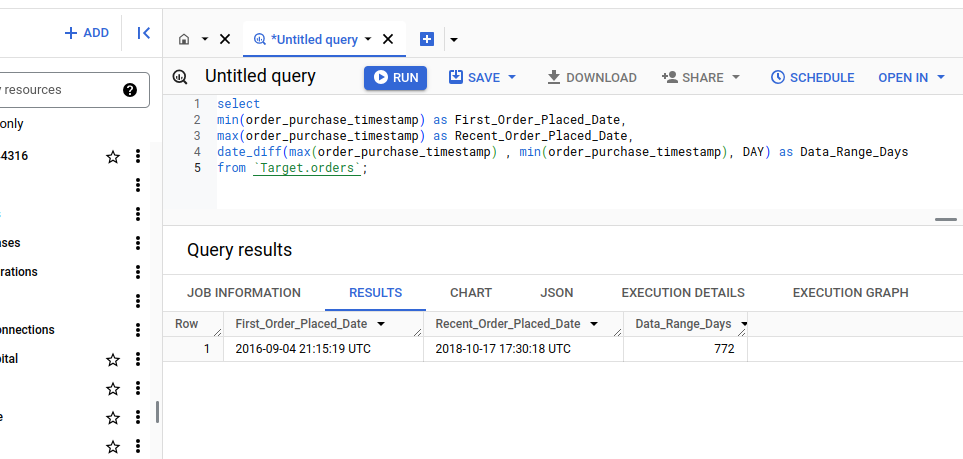
* Google BigQuery provides a convenient way to check all columns of a table & their data type.
* Click on Custome Table & Big Query will open a tab with Table Schema Details.
* Another way is to use *Information\_Schema.Columns*
* ***We observe that there are total 5 columns, 4 have string data type while 1 has timestamp***



### 

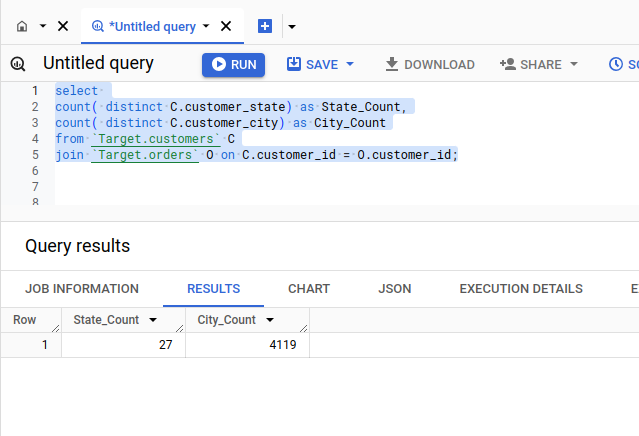
### Get the time range between which the orders were placed

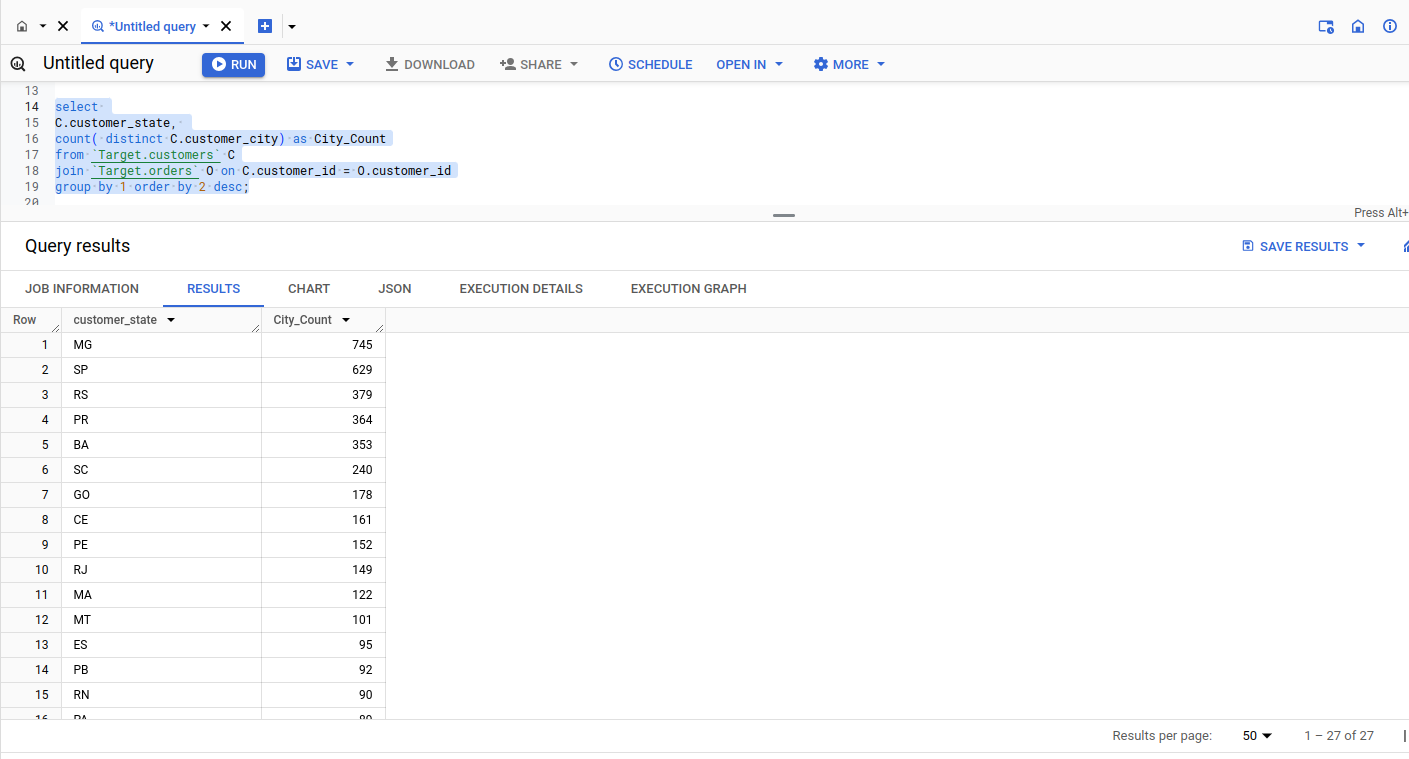
* We are looking for Oldest & Recent order datetime, Calculating difference between both of them as well.
* Since the same record is updated in the orders table, we do not filter on order\_status
* ***We Observe that We have 772 days of Range.***



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

* If we inner join both *Customers* & *Orders* table, than we get only customers who has placed a order, & it will automatically fall into timeframe of Orders.
* Here also we can provide data in 2 ways
  + Individual count of distinct cities & states.
  + Count of cities per state
* ***We observe that all customer who ordered something belong to 27 different states which has 4119 distinct cities, We also identified number of cities per state.***

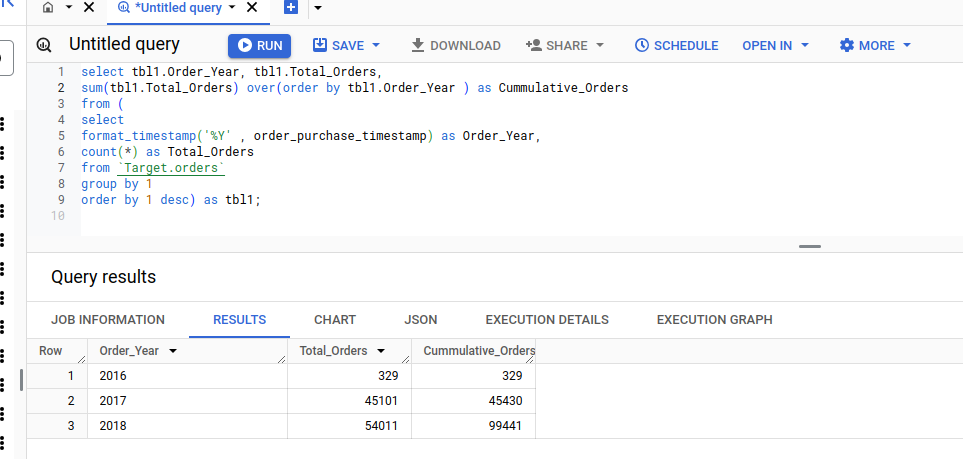


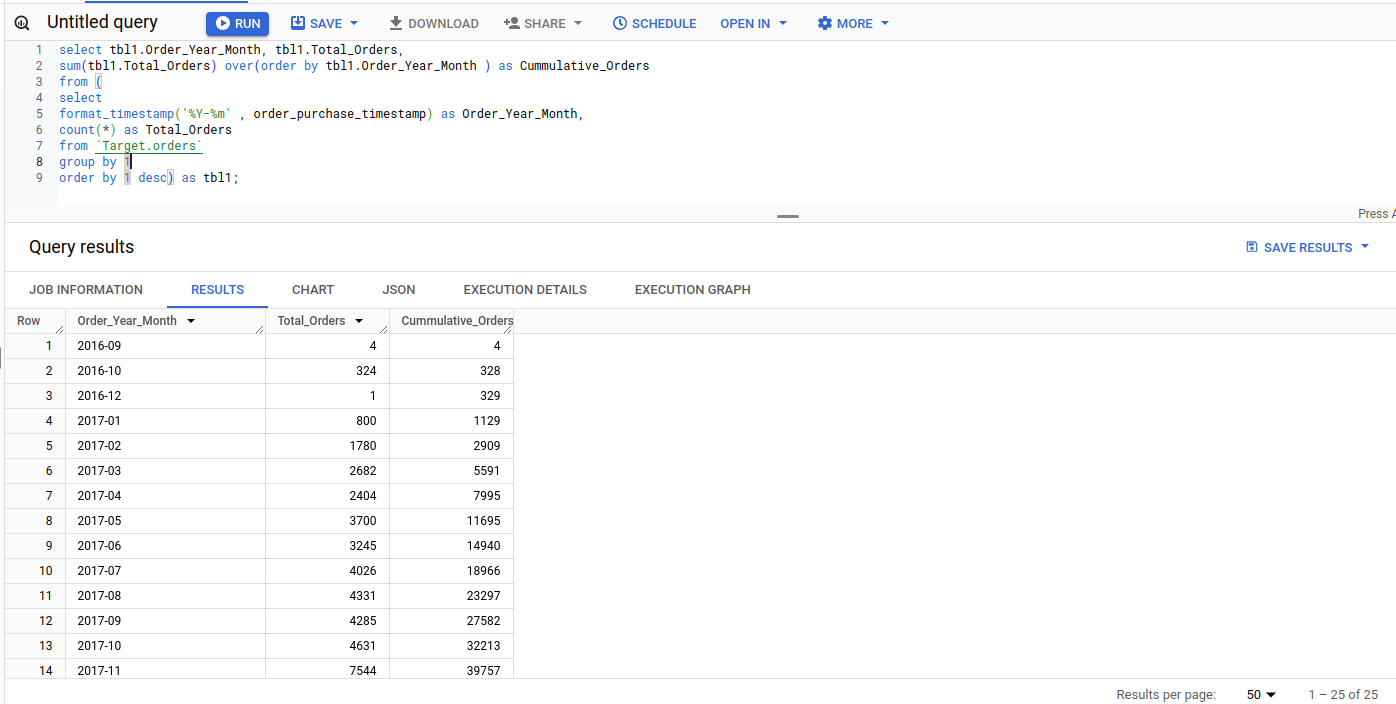


## In-depth Exploration

### Is there a growing trend in the no. of orders placed over the past years

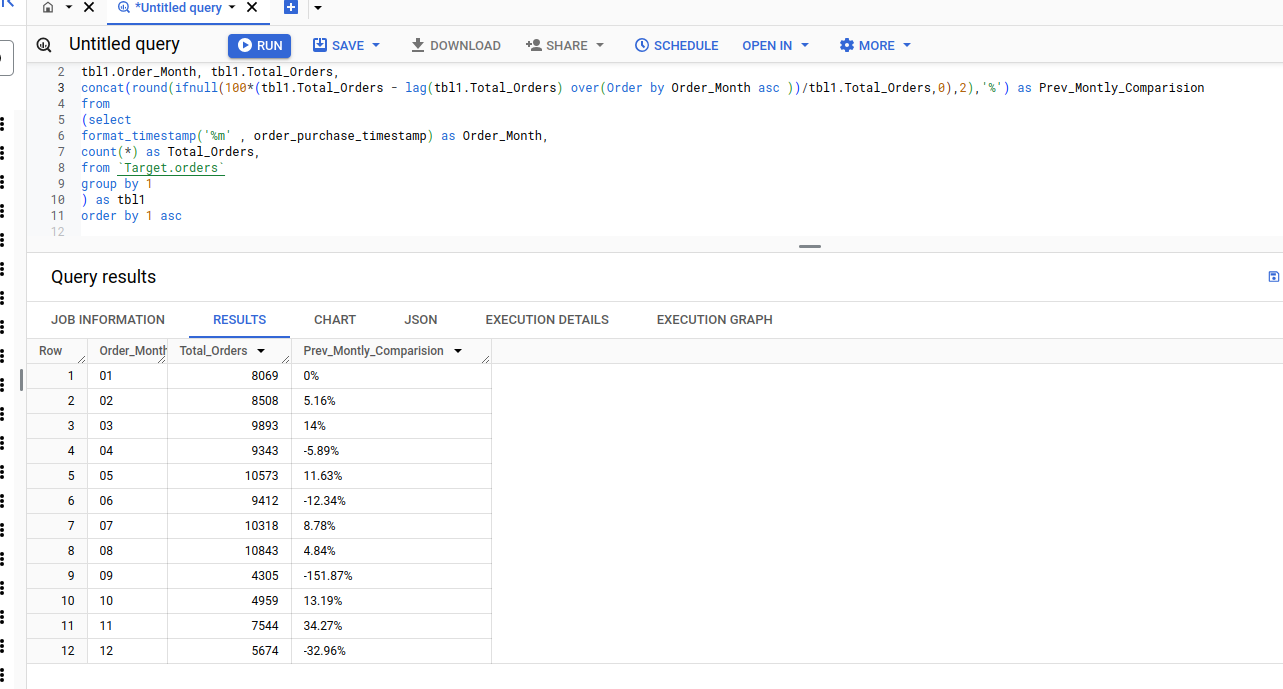
* To identify yearly tread we should get the number of orders *per year* or *per month per year.*
* Sort the Orders & calculate thier cumulative sum year on year
* Since we are considering trend of order placed , we should even consider canceled(625) and unavailable(609) orders
* ***We observe a Posotive Yearly Orders count pattern.***





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

* To identify monthly seasonality , We should group order bases on month.
* Since seasons ( e.g. winter/summer ) are based on months only , we should not consider Year-Month combination.
* Again we are considering all order irrespective of order\_status
* ***We observe that Monthly Seasonality***
  + ***Months May, June, July & August are best in terms of orders placed.***
  + ***Month Sepember, October, November & December are worst in terms of orders placed.***
  + ***Orders decline to lowest in November, While we receive most orders in August.***



### During what time of the day, do the Brazilian customers mostly place their orders

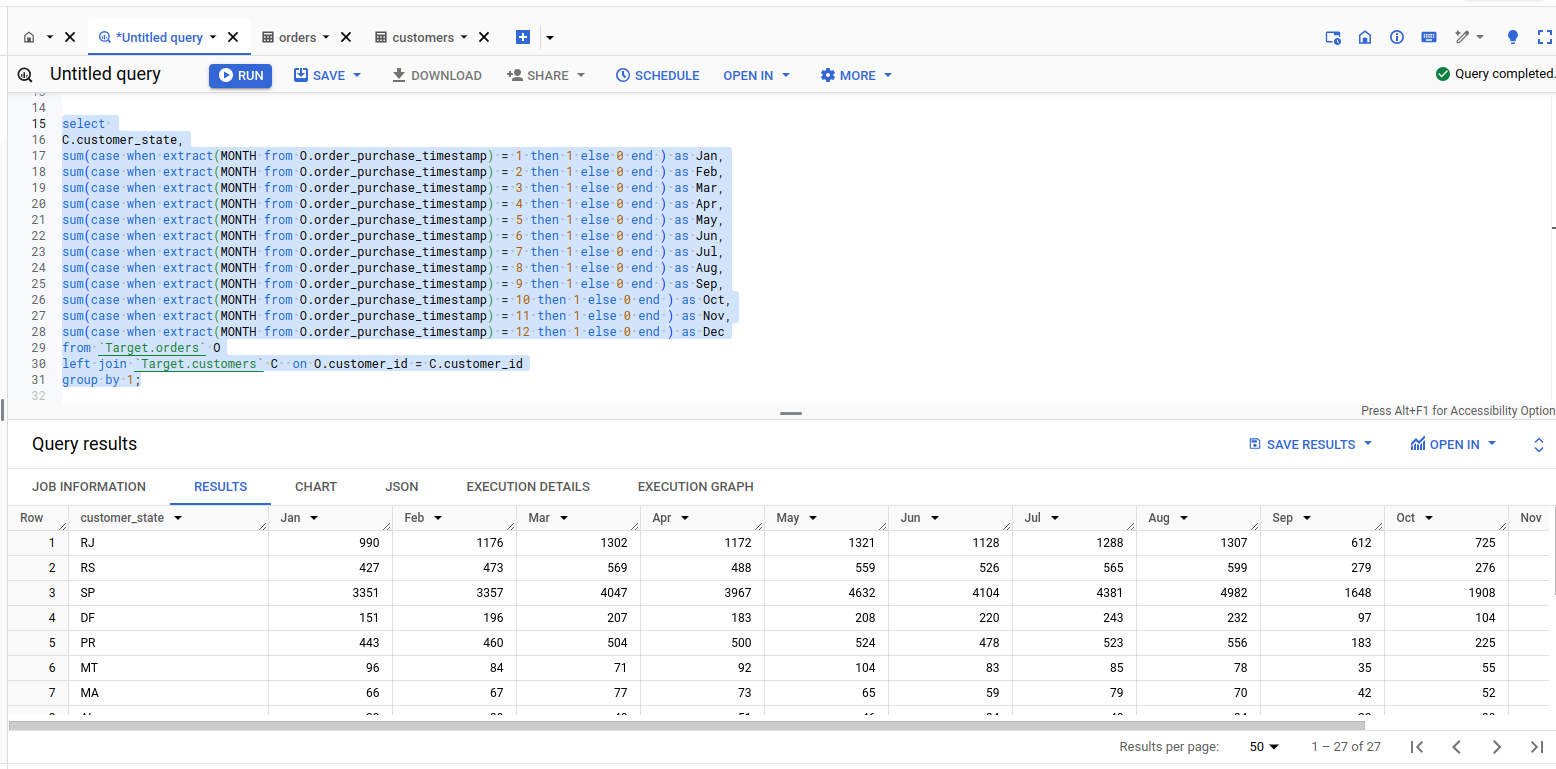
* Since the whole data is for Brazilian customer, There is no need to identify Brazilian Customer.
* First we need to extract the Hour part form the Order\_Purchase\_Timestamp.
* Next, Create categories with Case statement with given range of hours.
* Next, We count order placed in each category.
* As our final output requirement is 1 row, we limit 1 rows after sorting data based on order placed in each category in descending order.
* ***We observe that Brazilian customer mostly place thier orders in Afternoon.***

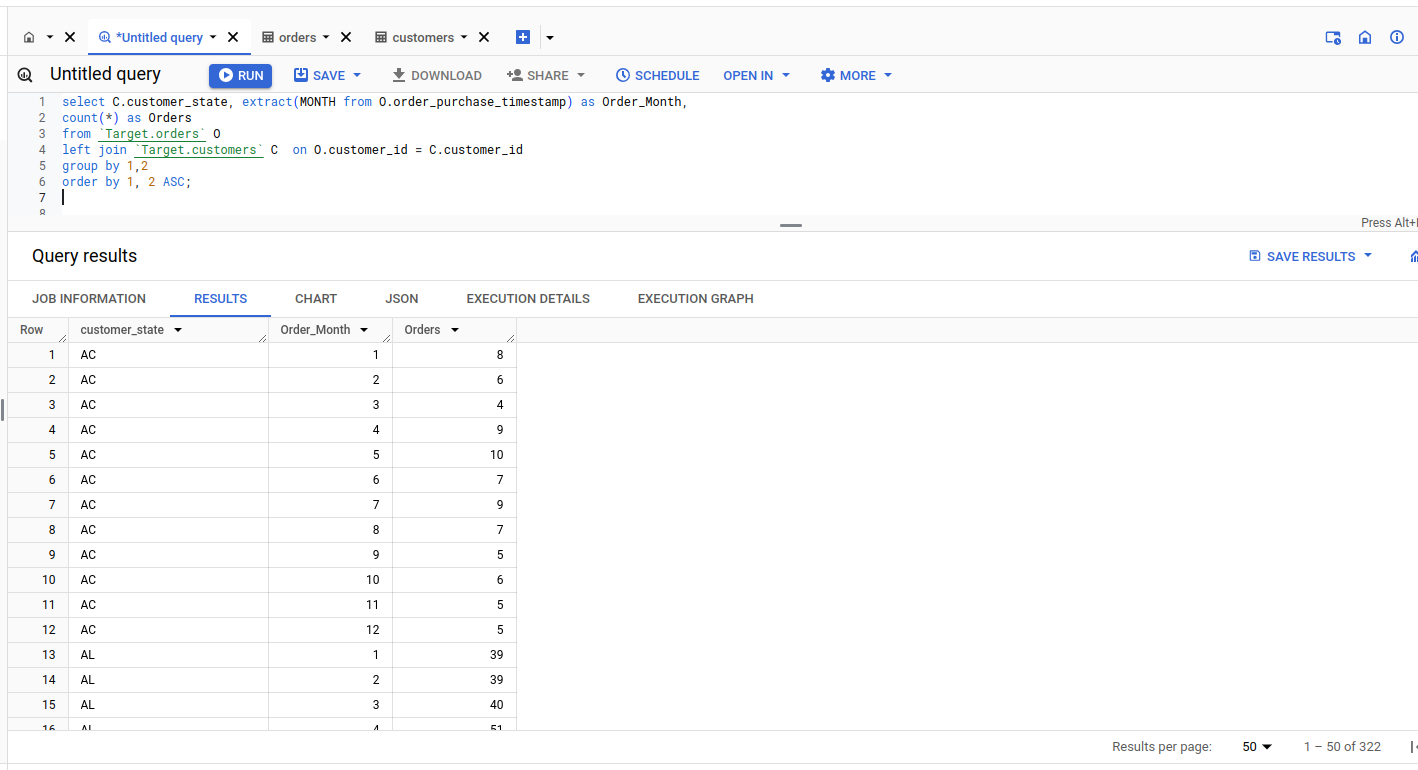


## Evolution of E-commerce orders in the Brazil region

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

* Orders table do not have *State* column.
* We should *left join* with *customers* table , so that we get State details, also retain data for orders.
* We can work with null State column separately. (*Curretly we dont have such data.)*
* We can represent this data in two format.
* ***We are able to observe the month on month orders placed in each state, however First solution provides a good pivotal view.***



**

### How are the customers distributed across all the states

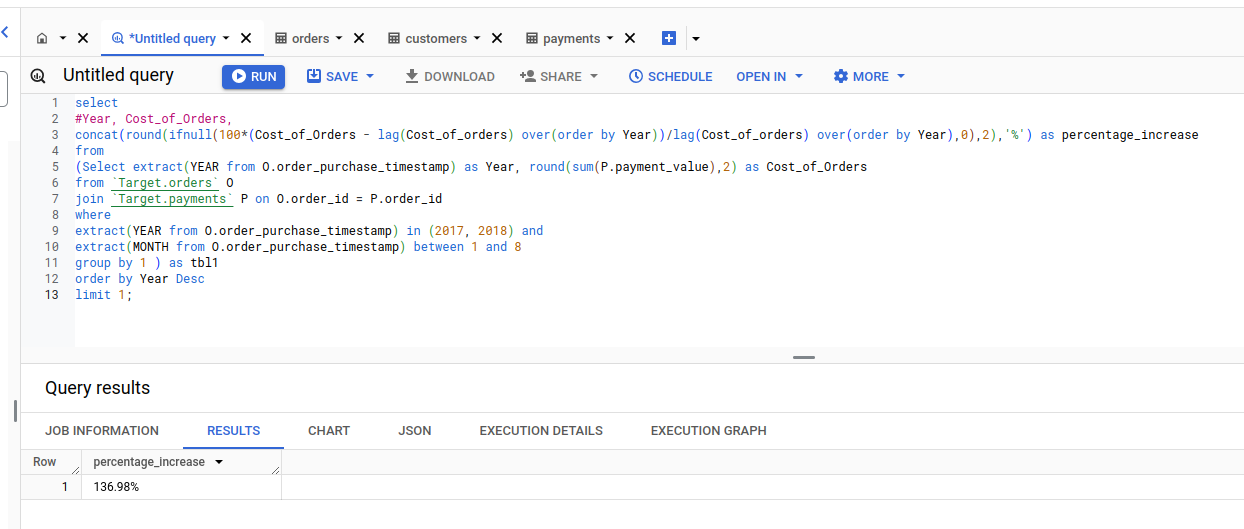
* There are two columns, *customer\_id* and *customer\_unique\_id,* But upon checking with *customers* & *orders*, we identify that each customer\_unique\_id has multiple customer\_id which is directly related to order\_id.
* Hence taking the count of customer\_unique\_id group by will provide the correct result.
* ***We Observe Customer distribution across state. SP , RJ & MG are top 3 contributing states.***

## 

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

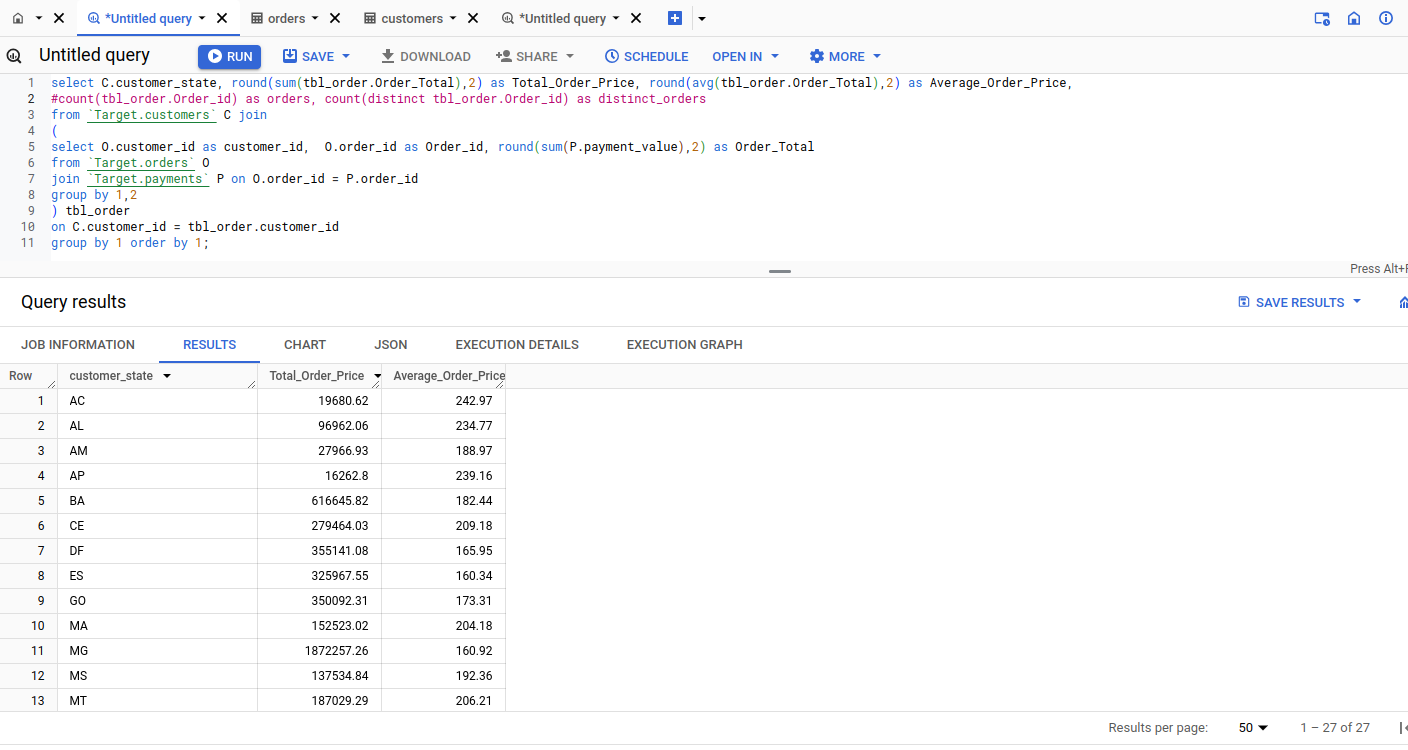
### Get the % increase in the cost of orders from year 2017 to 2018 (include months between Jan to Aug only)

* We identify that Orders & Purchase table has 1 to many relationshiop.
* Purchase table do not have any date column, we can get purchase date details by combining the Orders & Purchase table over order\_id.
* We aggregate the purchase\_amount over given year & month commbination in a inner query.
* We use lag function to get next value & calculate differnce. Since we are dealing with only 2 year, One of the output for lag will be null.
* We limit our Rows to only 1 to showcase percentage increase between 2017 & 2018.
* ***We observe Percentage Increase for Cost of Orders , Between January to Auguest months of 2017 & 2018 is almost 138%.***

******

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

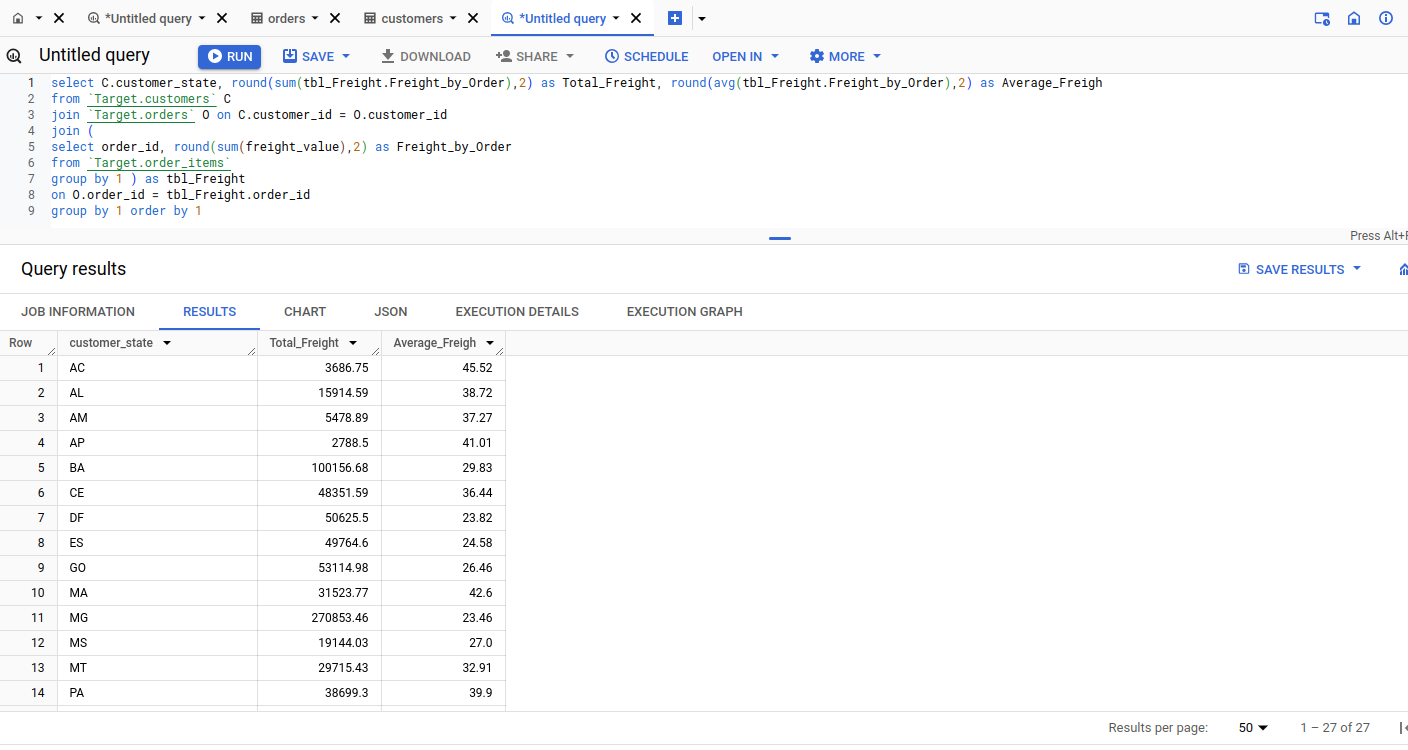
* To get order price statewirse, we need to join Payment, Orders & Customers.
* Since one order can have multiple payment, Hence calculating average using payments table directly is not a good options.
* First we create a inner query to calculate Order value per order. Combine it with Statedata & then calculate Statewise Total & Average.
* ***We observe total & average order price for each state.***



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

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

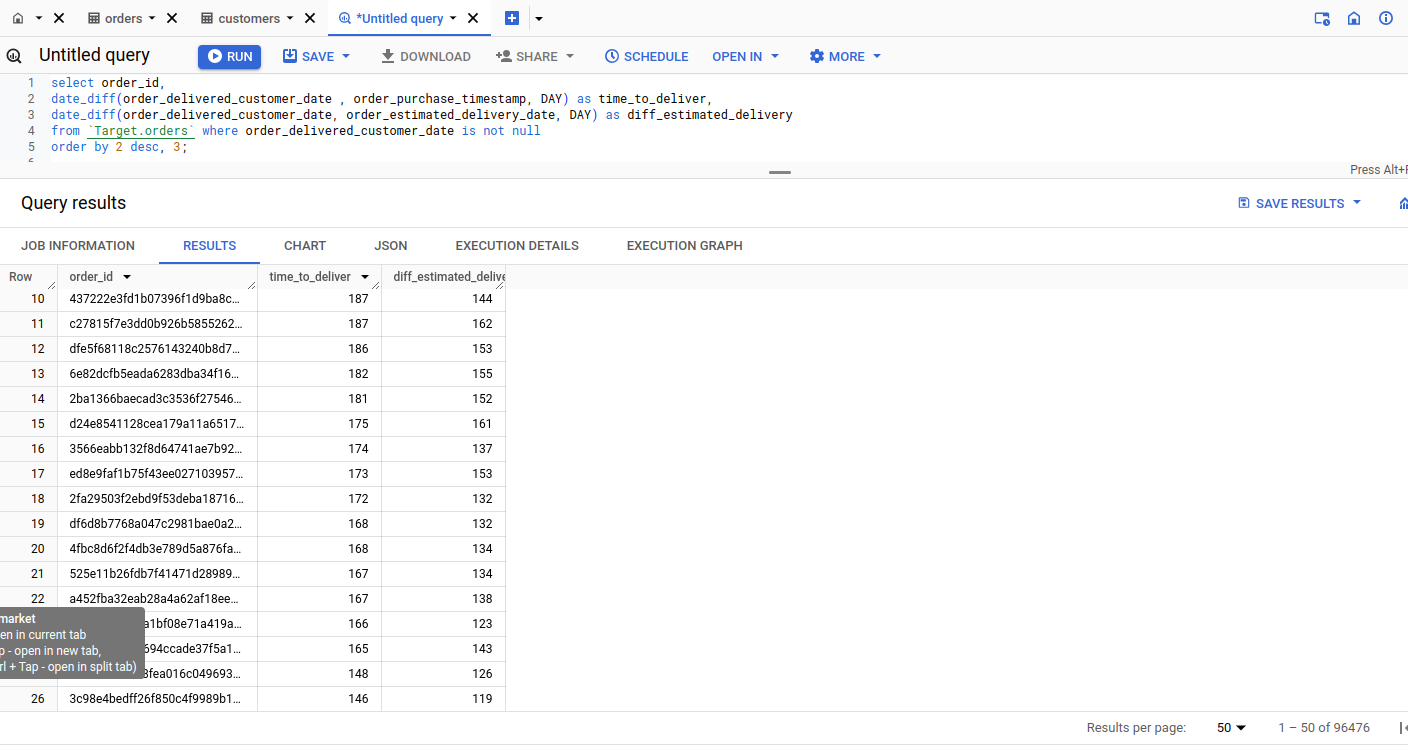
* Freight\_value is presnet in order\_items table, which has multiple rows for a single order, So we should first create a total freight\_value for each order.
* We further join it with orders & customer table to get to State for each order.
* Finally we can aggregate required values at state level.
* ***We observe total & average order freight value per state.***



## Analysis based on sales, freight and delivery time

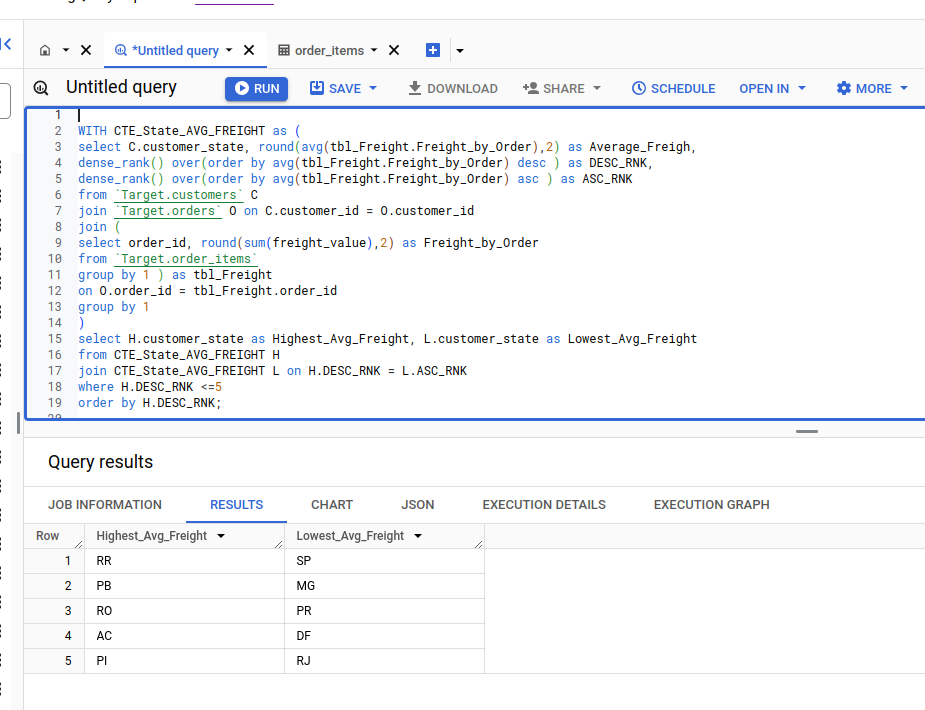
### Find the no. of days taken to deliver each order from the order’s purchase date as delivery time. Also, calculate the difference (in days) between the estimated & actual delivery date of an order.

* Intial inspection of *Orders* table suggest that customer delivery date is only applicable for order with either *delivered* or *canceled* status.
* We are considering *6 canceled* order which has customer delivery date populated for analysis.
* ***We observe that a lot of the orders has been delivered very late.***

******

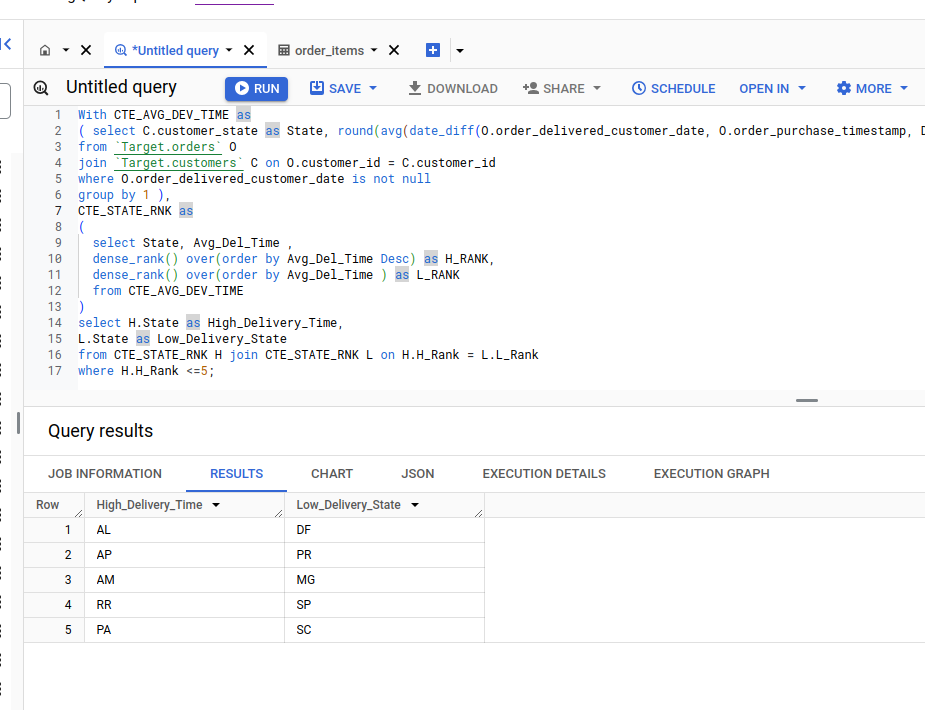
### Find out the top 5 states with the highest & lowest average freight value

* We had already calculated statewise average freight ealrier, we can extend same query to identify top 5 & low 5.
* However since we want to display both in adjacement column, we can first calculate state rank over average freight value in both orders & self join to display data in adjancemt columns.
* ***We observe Top 5 & Low states as per average frieght rate.***

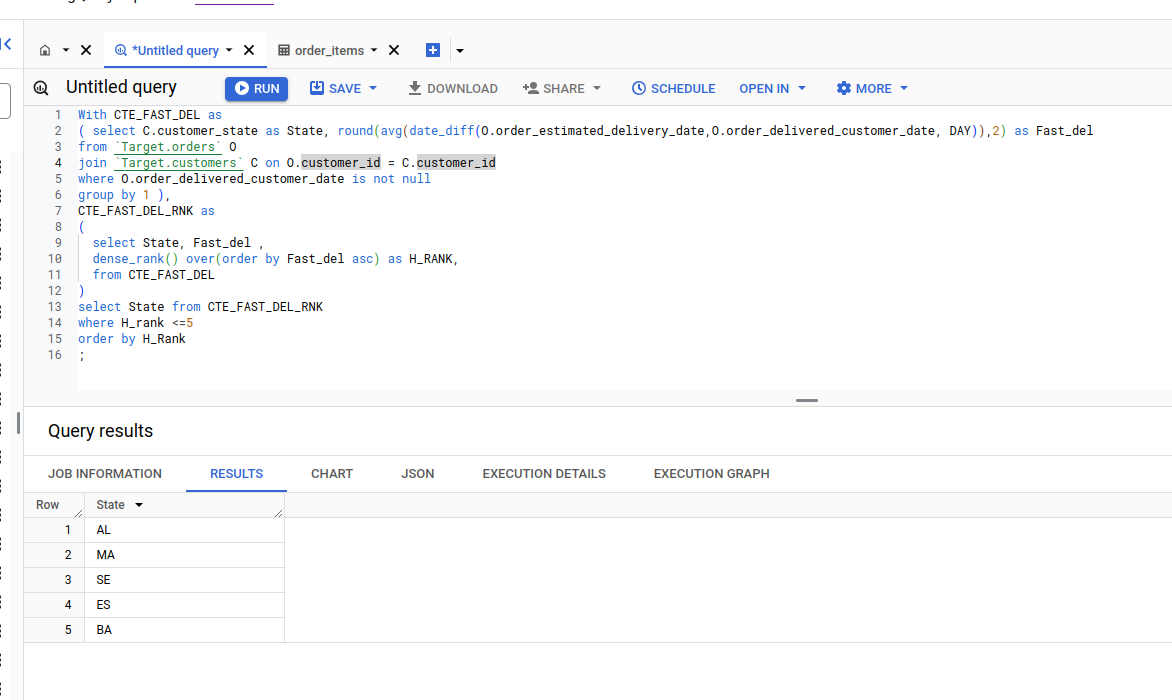
******

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

* We use the same logic that we have used for average freight value per state.
* ***We observe top 5 & low 5 state with average delivery time.***

******

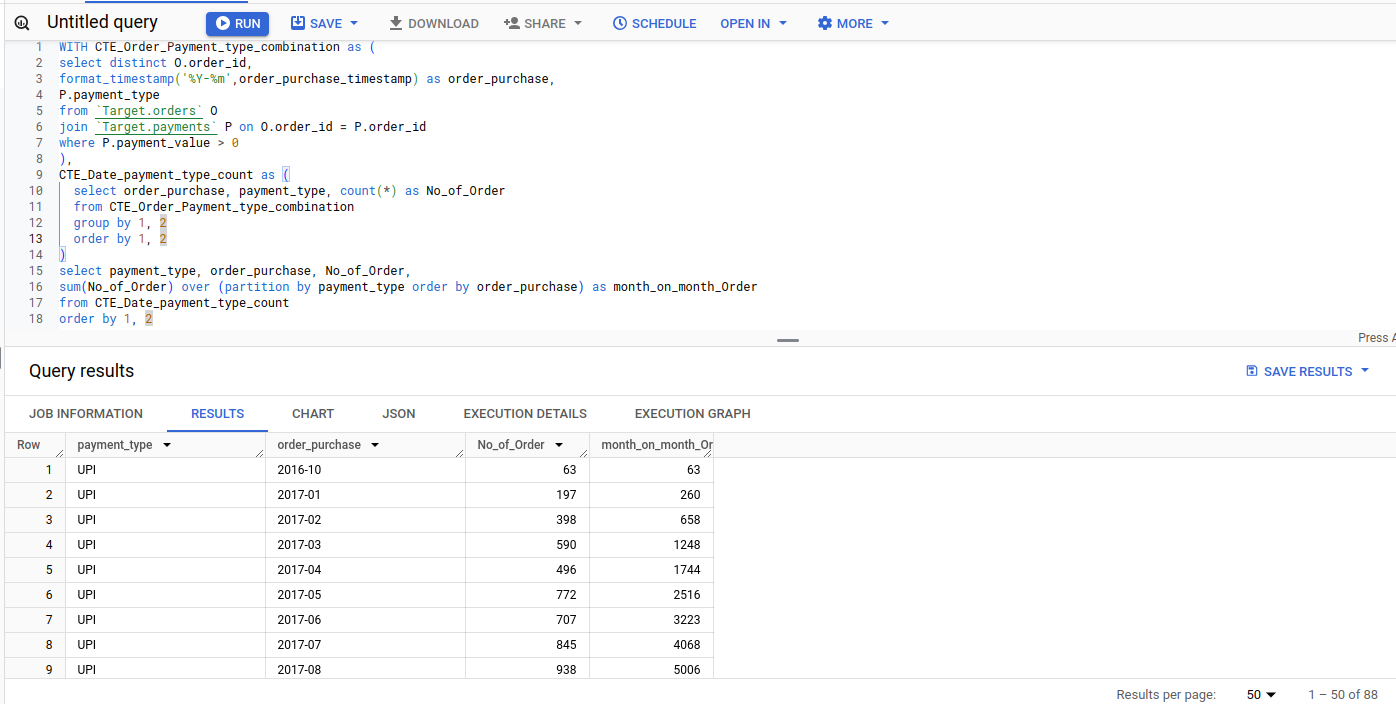
### Find out the top 5 states where the order delivery is really fast as compared to the estimated date of delivery.



## Analysis based on the payments

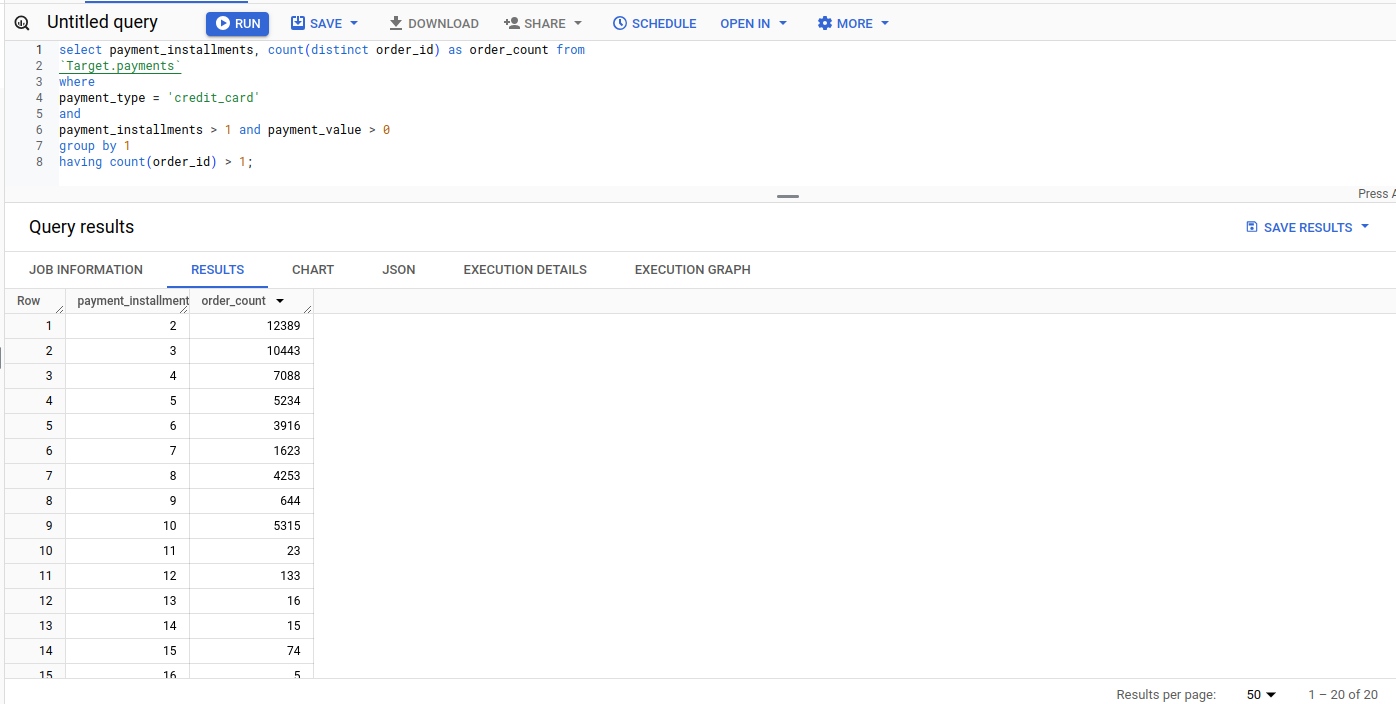
### Find the month on month no. of orders placed using different payment types.

* We analyse that one order can have multiple payments & payment type.
* We cnsider unique combination of order\_id, Order\_date & payment\_type for further analysis.
* Order with multiple payment type ( card & voucher ) will be counted in both payment type.
* We are consider only order where payment\_value > 0
* ***We Observe positive month on month trend.***

******

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

* Order where installments are opted by customer has payment type as *Credit Card*.
* We need to only include playment that has *payment\_value > 0, Payment\_installments > 1*
* *Atleast 2 payment* enries are present.
* ***We observe orders on installement where atleast 1 emi has been paid.***

**

## Actionable Recommendation

1) Target should try to increase city center in state where overall city center count is less than 100. 12 State has more than 100 city centers while 15 State has less than 100 city centers.

2) Target should try to identify cause of rapid drop in customer order starting from September till December.

3) Most Suitable time for Ad compaign & new launches seems be between May & August.

4) Since Orders placed in Dawn period are very few (5.27%), It could be utilized for Production issue fixes & maintenance activities.

5) Target should get rid of Customer\_Id column in Customer table, With Each new order a New Customer\_Id is generated which is tagged to Customer\_Unique\_id. Because of this the size of customer table will continue to grow parallel to new orders. Unless its fulfilling some other purpose which we have not discovered based on this Case Study.

6) Target should try to increase customer base in state where customer count is less than 2000, Currenlty 19 State cout of 27 have less than 2000 customers.

7) Target should improvise on delivering the orders faster, 63 order out of 96476 were delivered in over 100 days.

8) Target should improve on providing the estimated delivery date, Only 2754 orders were delivery on estimated delivery date,

9) Target should improve average delivery time against expected delivery time in 5 states where it is more than 15 days.