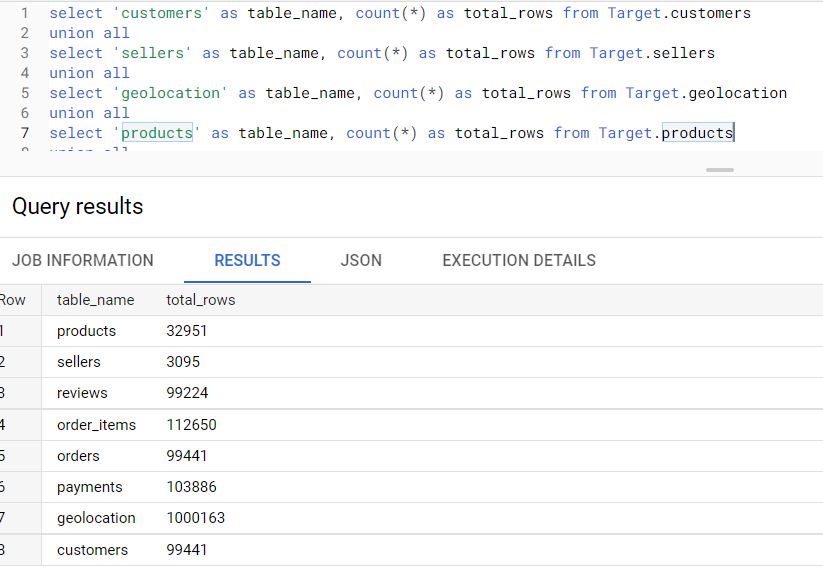
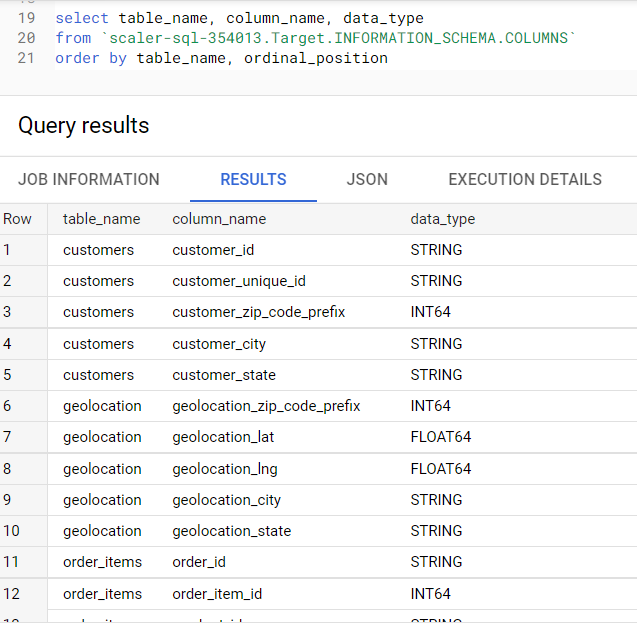
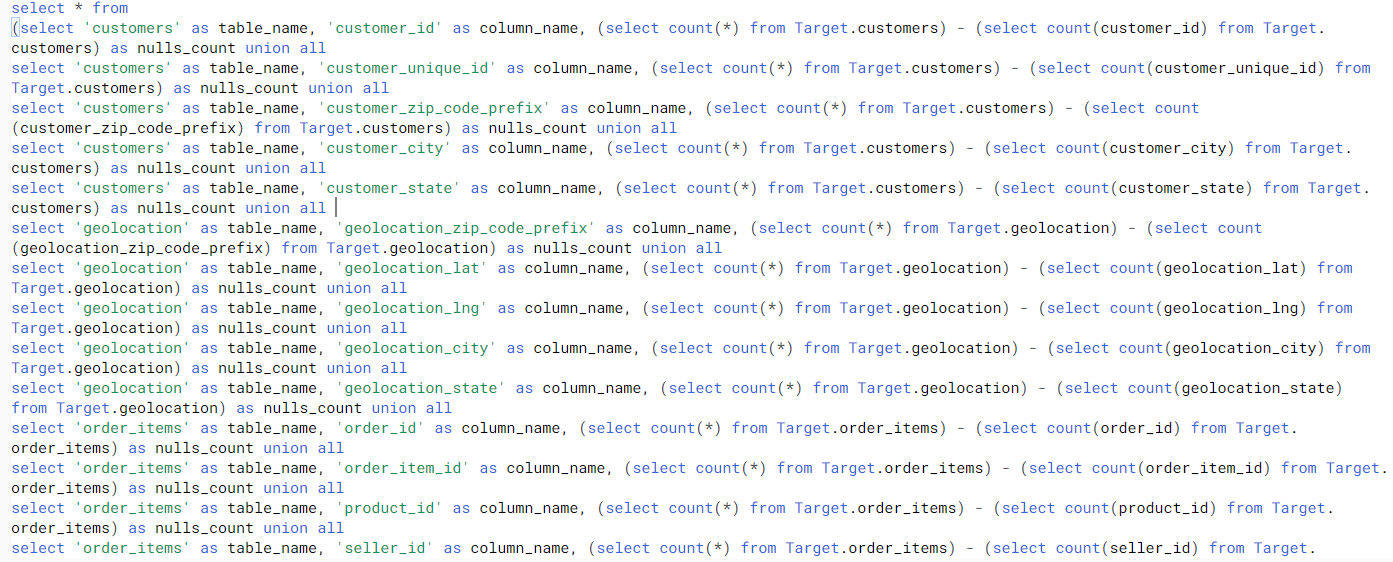
1. **Import the dataset and do usual exploratory analysis steps like checking the structure & characteristics of the dataset.**
2. Count of each table: This verifies that all the export to BigQuery has been done successfully as the record count in provided excel matches with the count of each table in BigQuery.

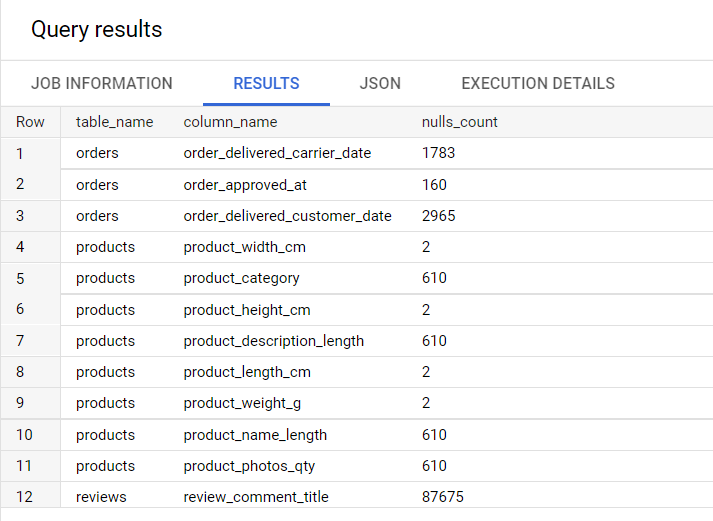


1. Data type of each column in each table: It shows that data types have been converted properly.

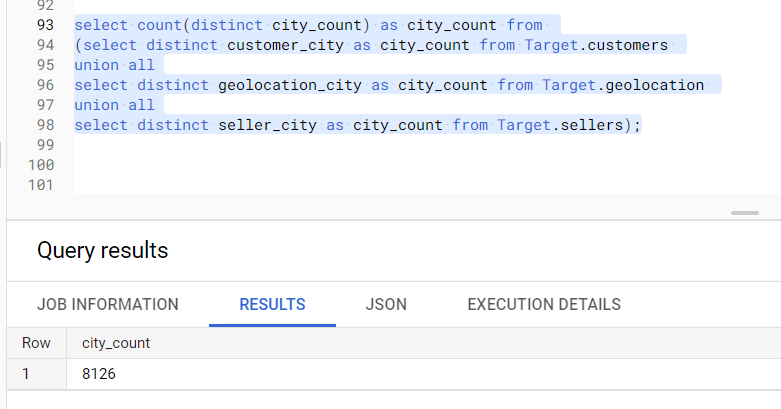


1. Missing values in each table:

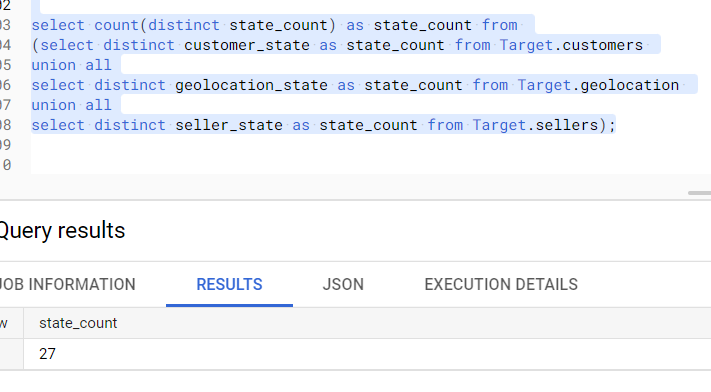




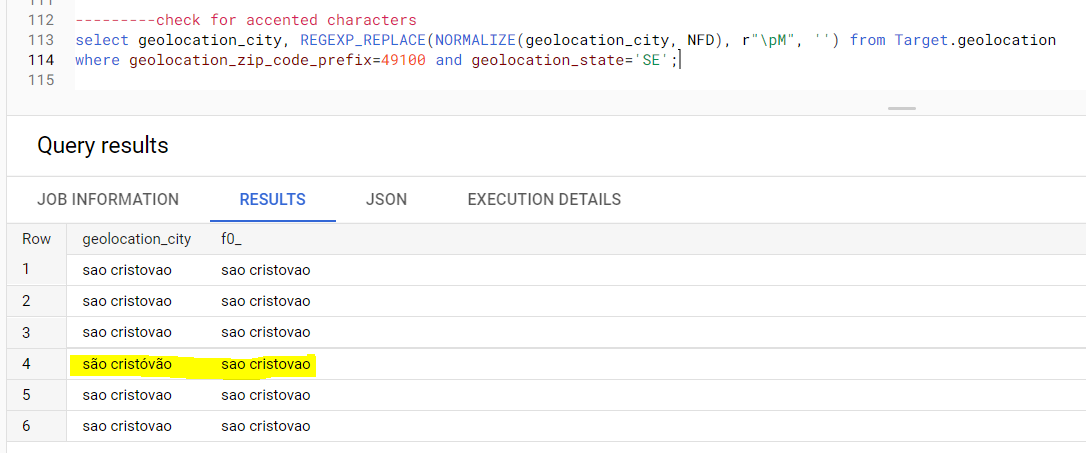
1. Total city count: There are total of 8126 different cities in Brazil.



1. Total state count: There are a total of 27 states in Brazil.



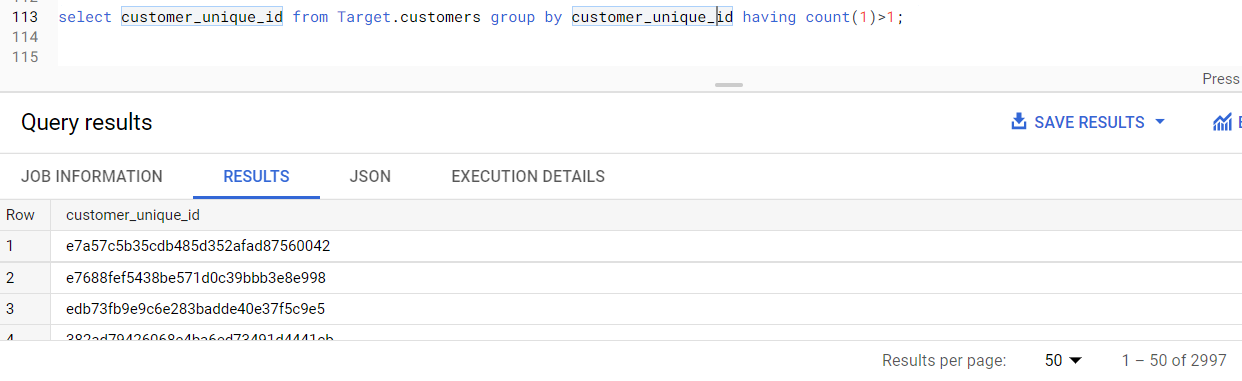
1. Now we see that there are text columns which has accented characters, it will create issue while joining with other tables. For this BugQuery provides Normalize function which can be used to convert accented characters to normal English alphabets.

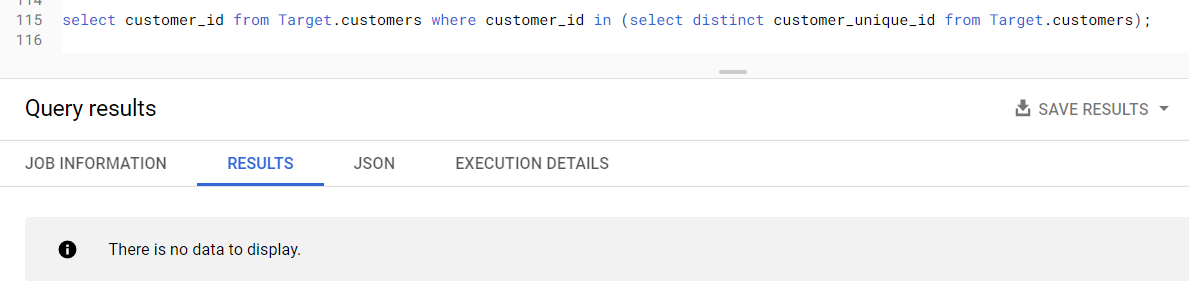


1. **Customer table Analysis: PK customer\_id**

This shows that customer\_unique\_id has 2997 duplicates. Customer\_unique\_id is associated to different customers, however customer\_id is just a unique id for each purchase.

This can also be seen below as customer\_unique\_id and customer\_id are mutually exclusive.





Below shows that some customers have move out to different places as their zip codes are different.



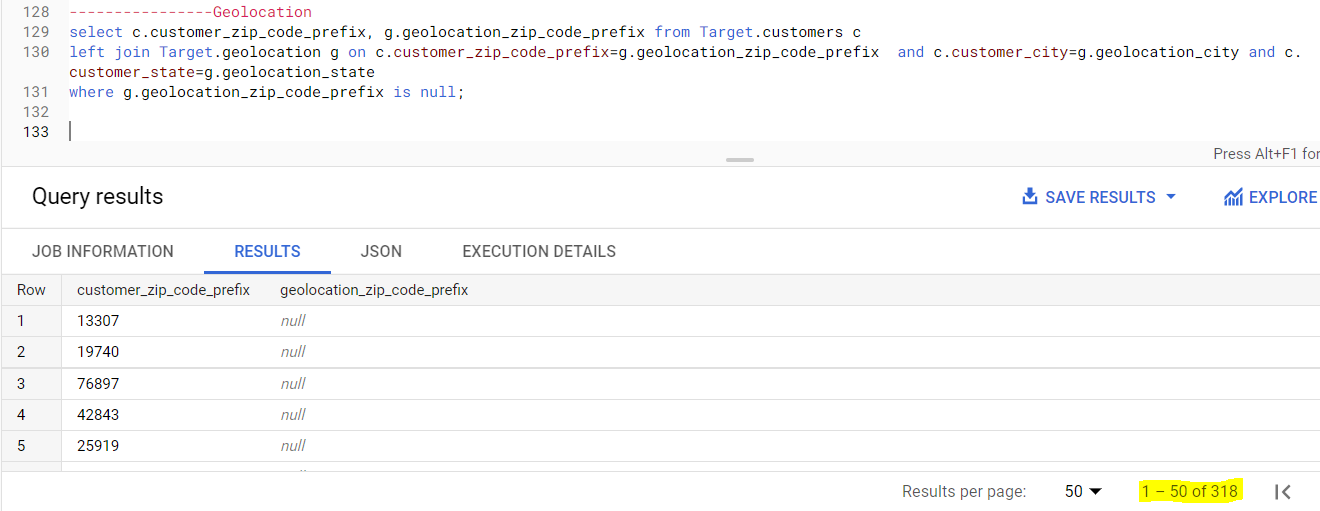
1. **Sellers table Analysis: PK seller\_id**

Unlike customers sellers have been staying in the same location.

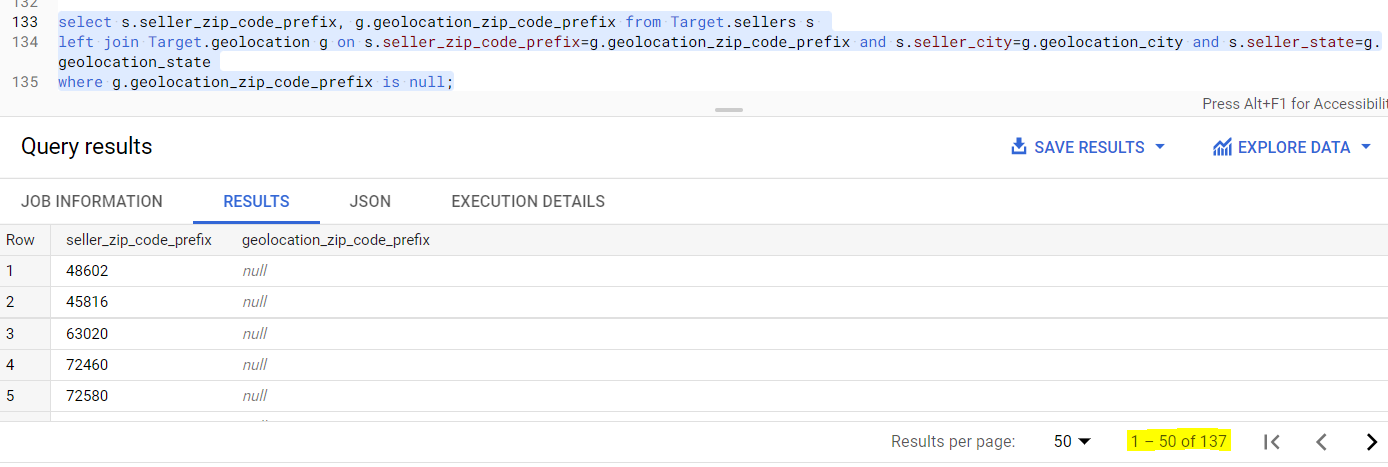


1. **Geolocation table Analysis: PK**

There are 318 geolocations zip codes that are not present in geolocations table but present in customer table.

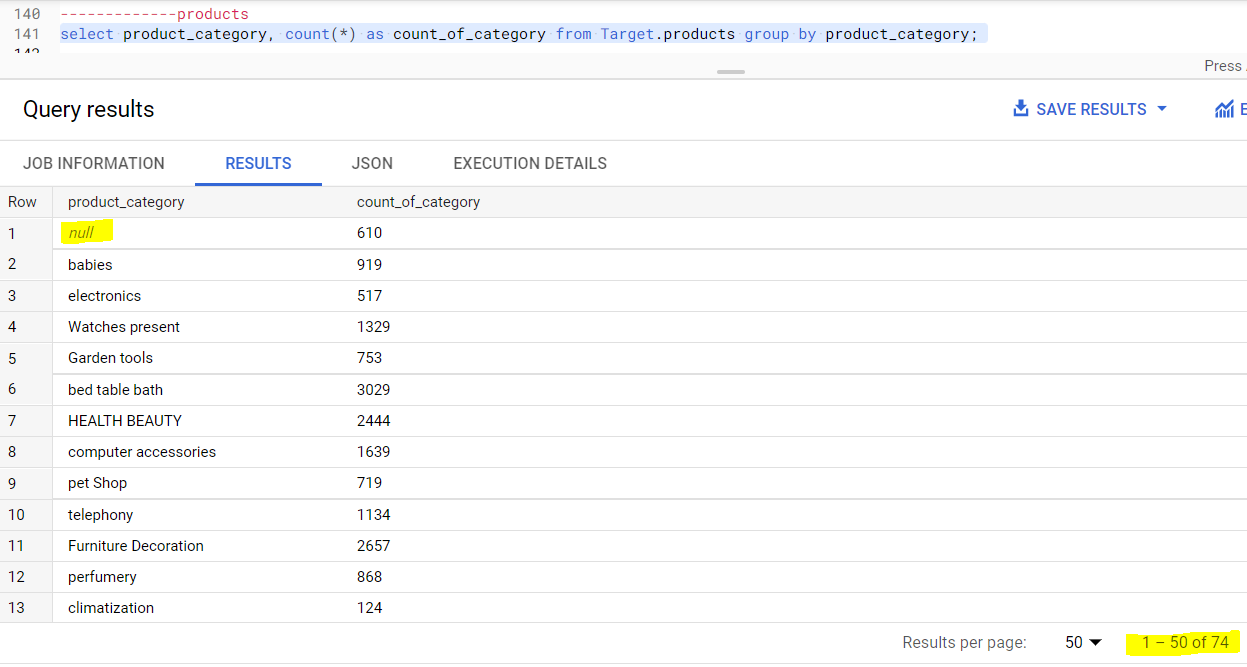


There are 137 geolocation zip codes that are not present in geolocation table but present in seller table.



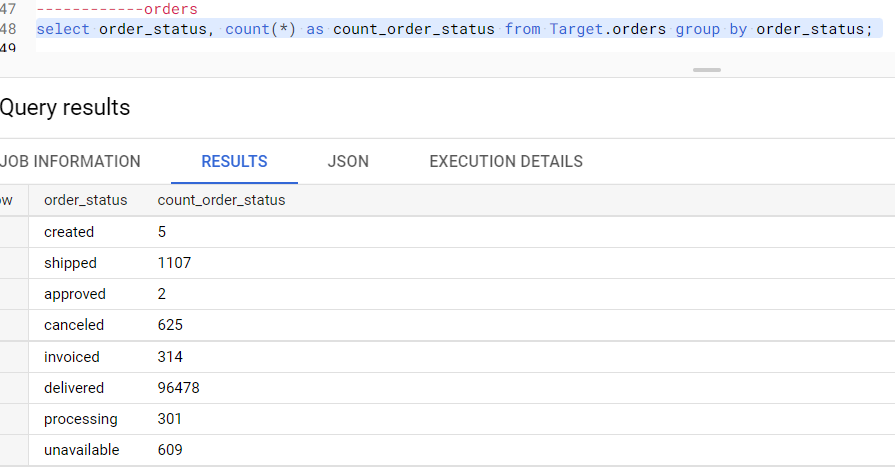
1. **Products table Analysis: PK product\_id**

There are 74 different types of product categories, also there are 610 products where product category is not present / NULL.

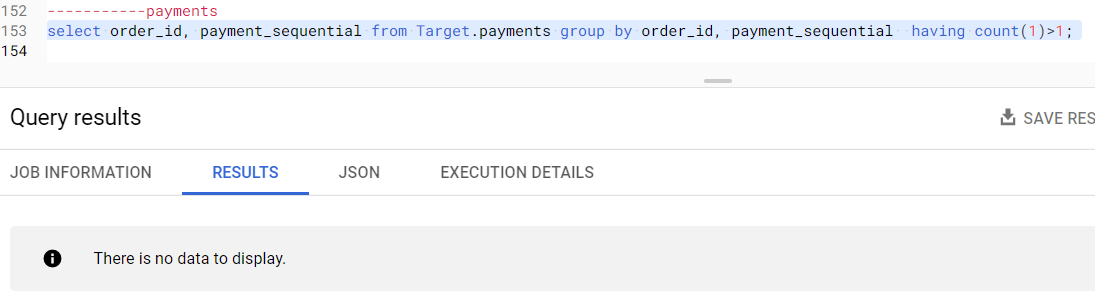


1. **Orders table Analysis: PK order\_id**

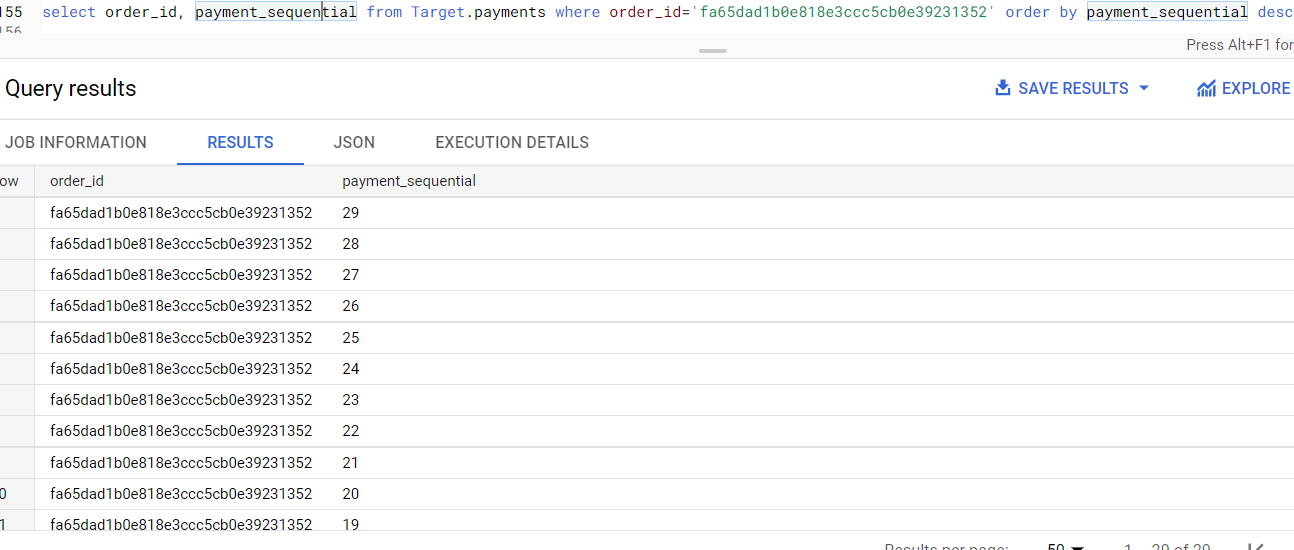
There are 8 different order status, most of them are delivered.



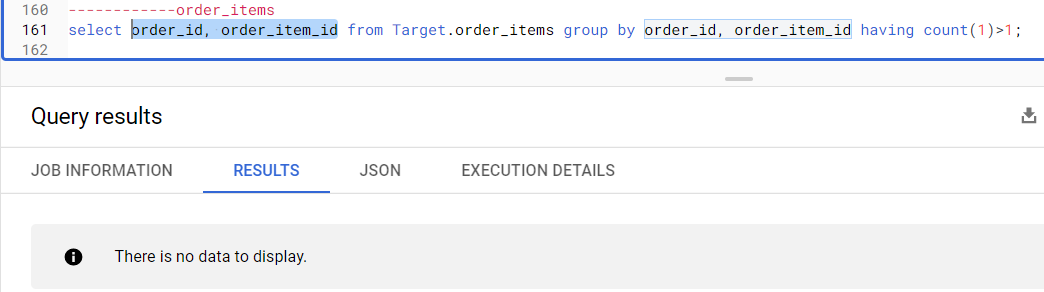
1. **Payments table Analysis: PK payment\_id, payment\_sequential**



We can see that payment sequential has number associated to each payment done in case of EMIs.



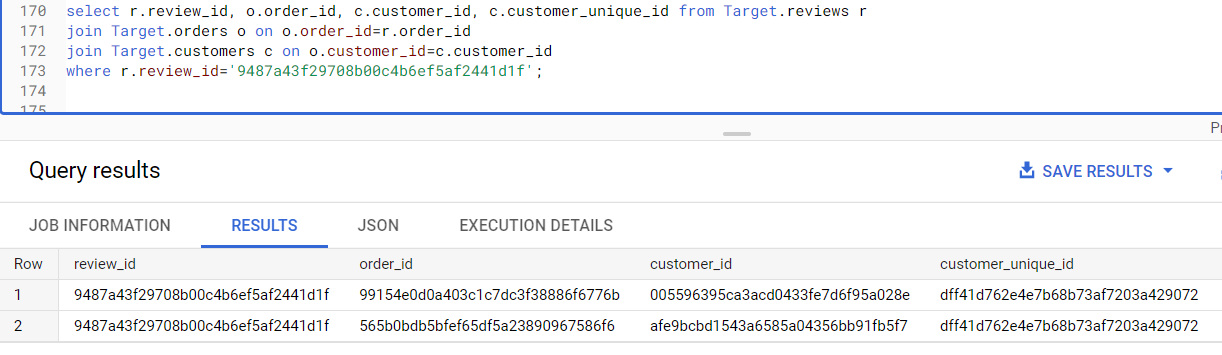
1. **Order Items table Analysis: PK order\_id, order\_item\_id**



1. **Review table Analysis: PK review\_id, order\_id**

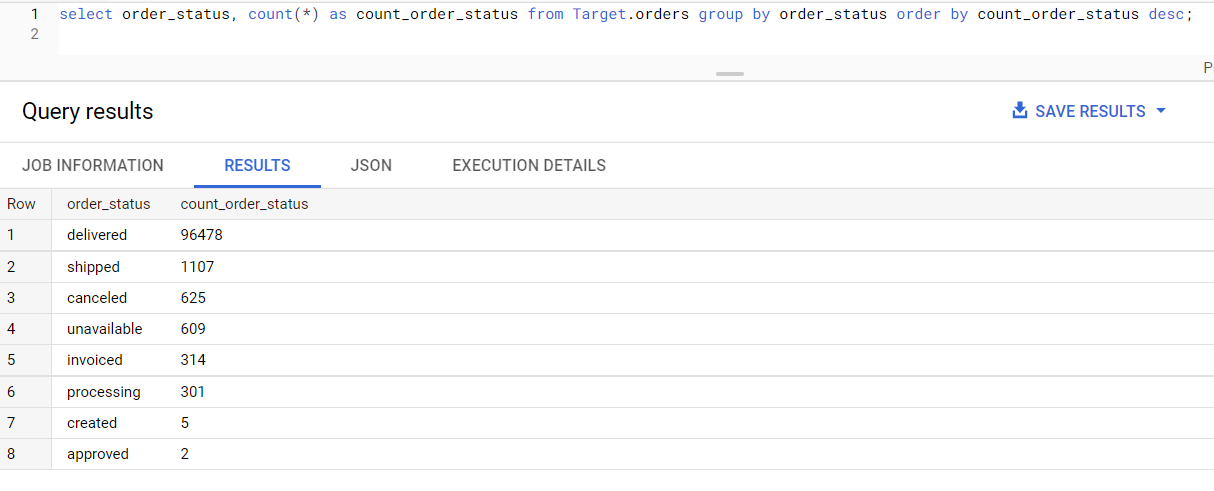


If we analyze one review which is having multiple order\_ids at first glance it feels like, how can two different orders have one review. But diving deep we found that those orders are done by a single customer. One example given below.



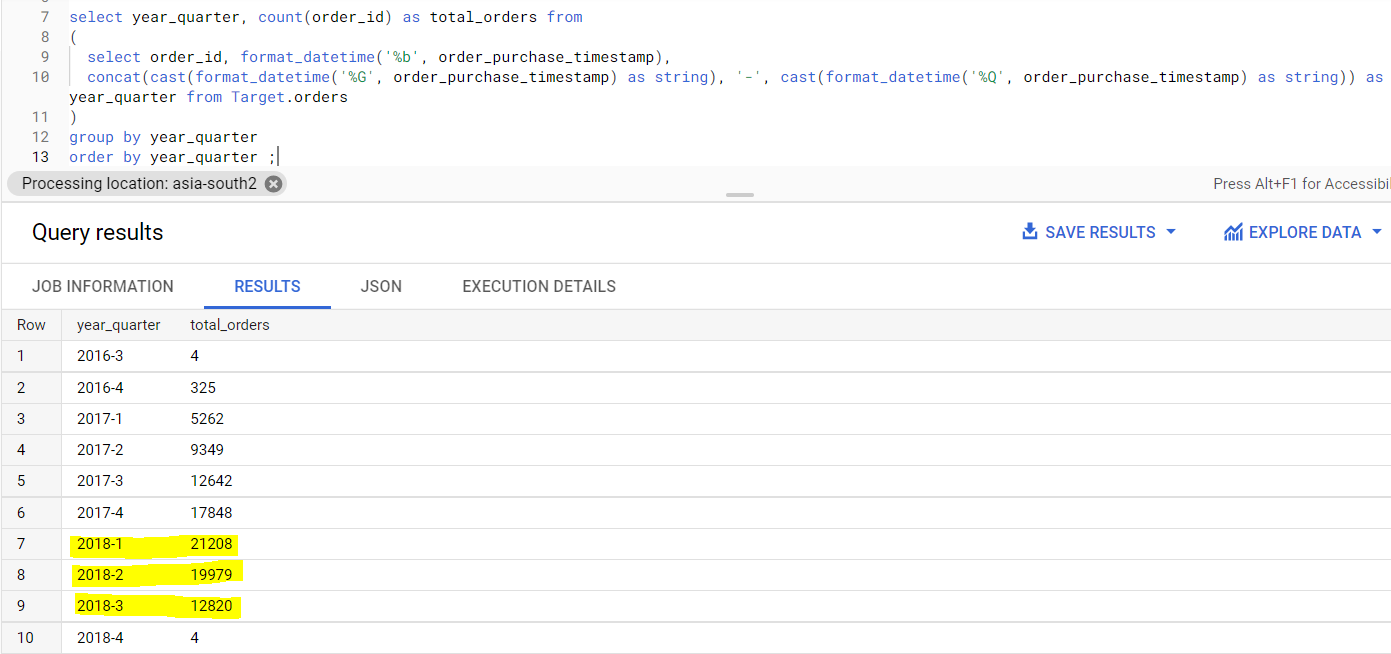
1. **In-depth Exploration:**
2. How many orders do we have for each order status?

We can see that most of the orders are delivered, some are shipped, canceled and so on.



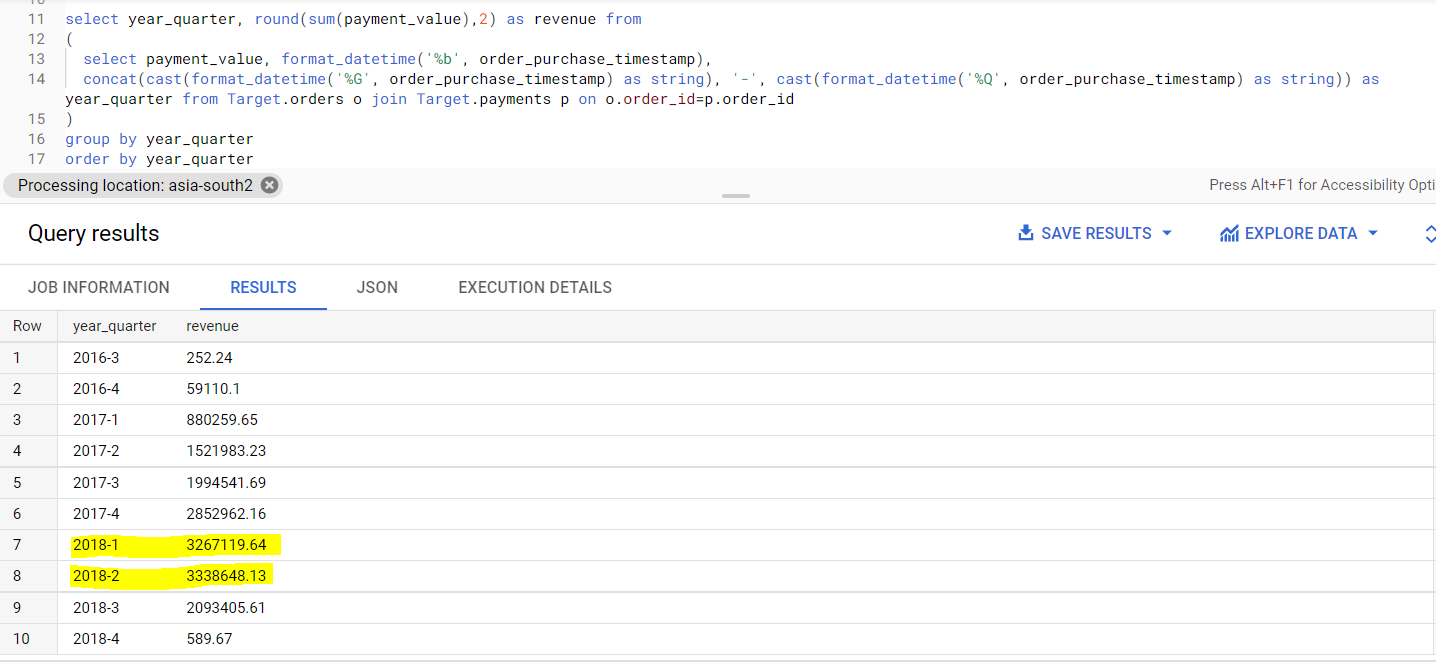
1. Is there a growing trend on e-commerce in Brazil? How can we describe a complete scenario?

If we check quarter wise number of orders we can see an increasing trend. Expect for 2018 quarter 2 decreased from quarter 1 and quarter 3 decreased from quarter 2.

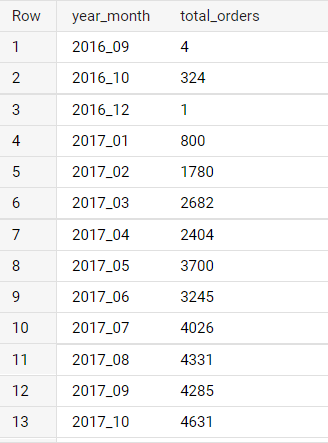
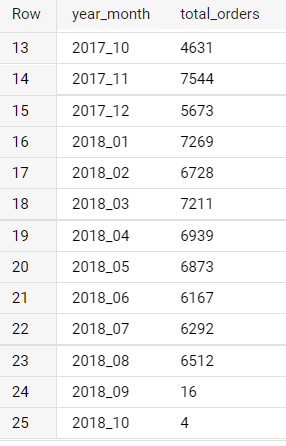


Now if we check for the revenue / sum of total payments received from customers then we see that though in the above graph the counts of orders from 2018-01 to 2018-02 has decreased by roughly 12000 still the total revenue has increased.

This shows that customers have been purchasing costlier products or they have been purchasing multiple products in one go.

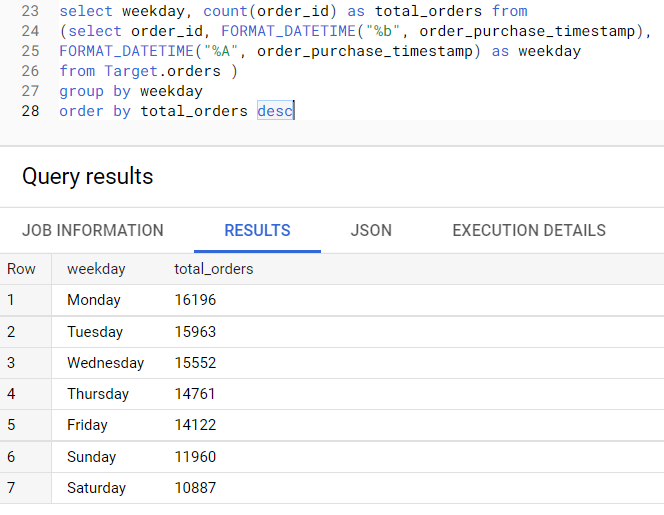


Similar way if we check for month wise orders there are some months which has less number of orders than the previous month, but overall it’s increasing.

1. Which day has the highest orders?

Interestingly the number of orders keeps on decreasing as the week progresses. Weekends has very less orders compared to Monday and Tuesday.



1. What time does Brazilians tens to buy?

Dawn – 0 to 6

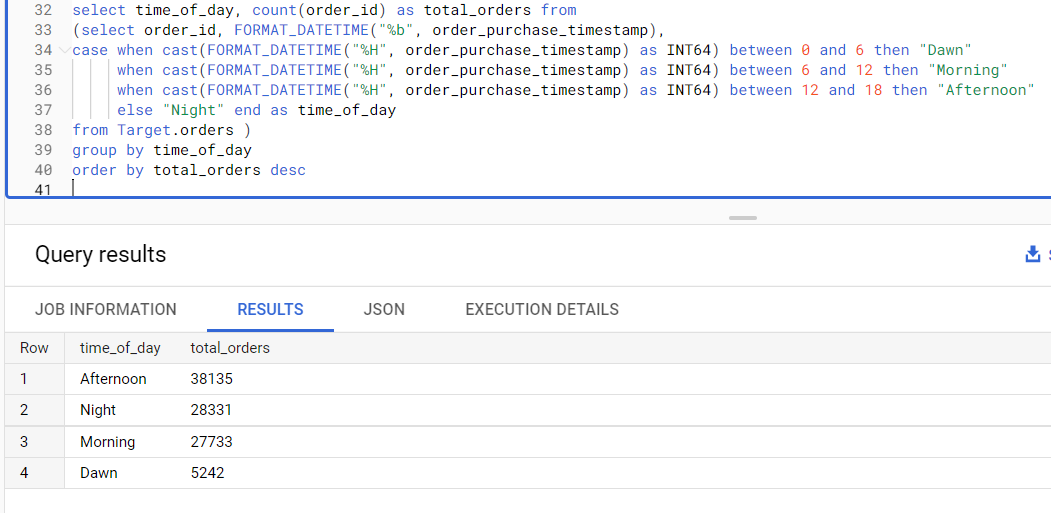
Night – 6 to 12

Morning – 12 to 18

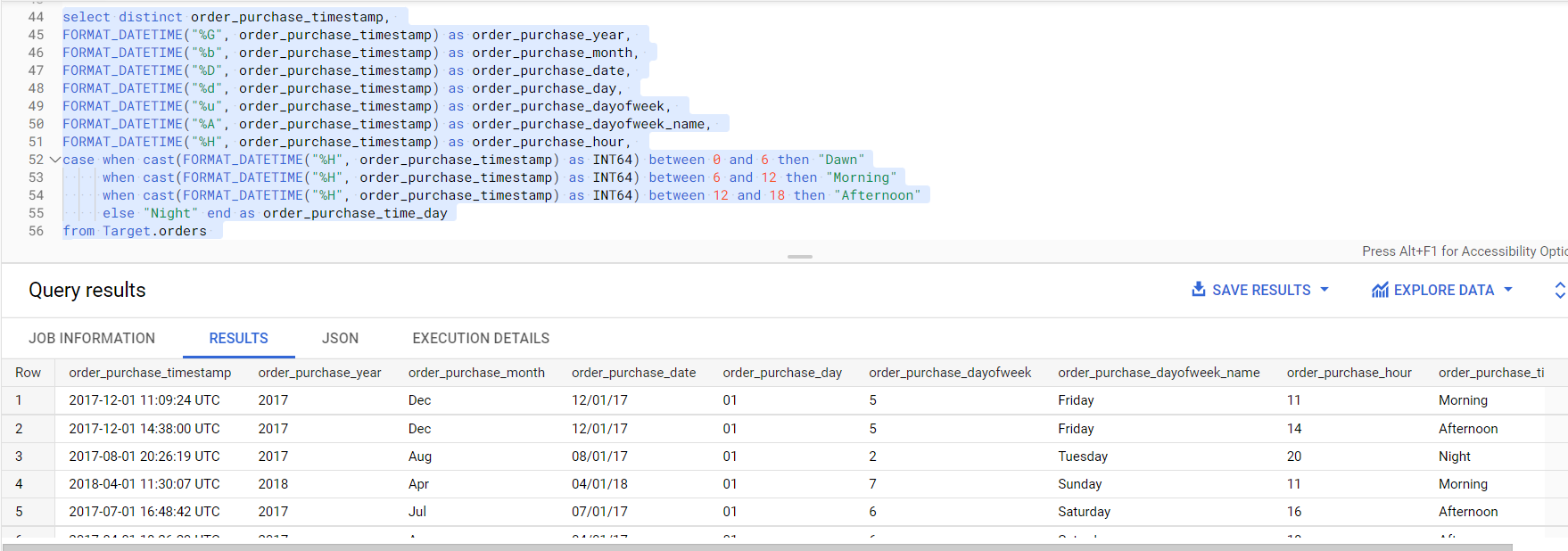
Afternoon – 18 to 24

We divide the day into 4 parts.

Afternoon has the most orders, morning and night are almost the same. But the Dawn has the least orders as expected.



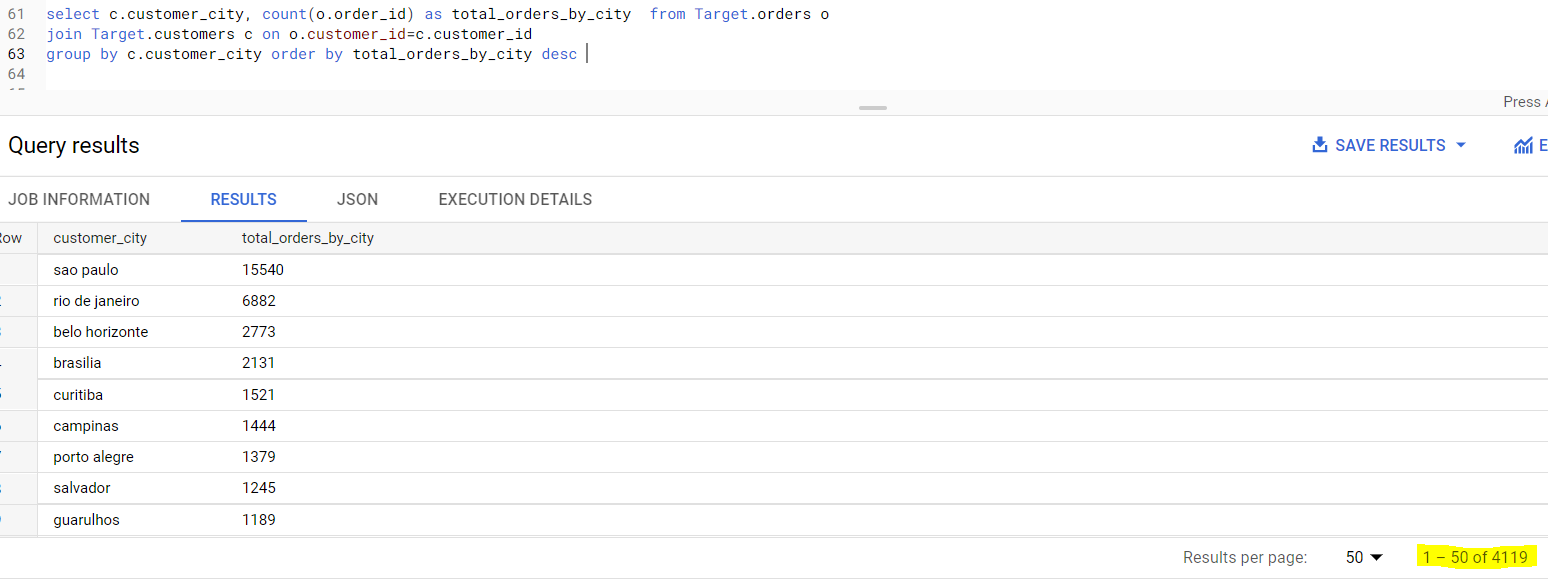
1. Feature Extraction:



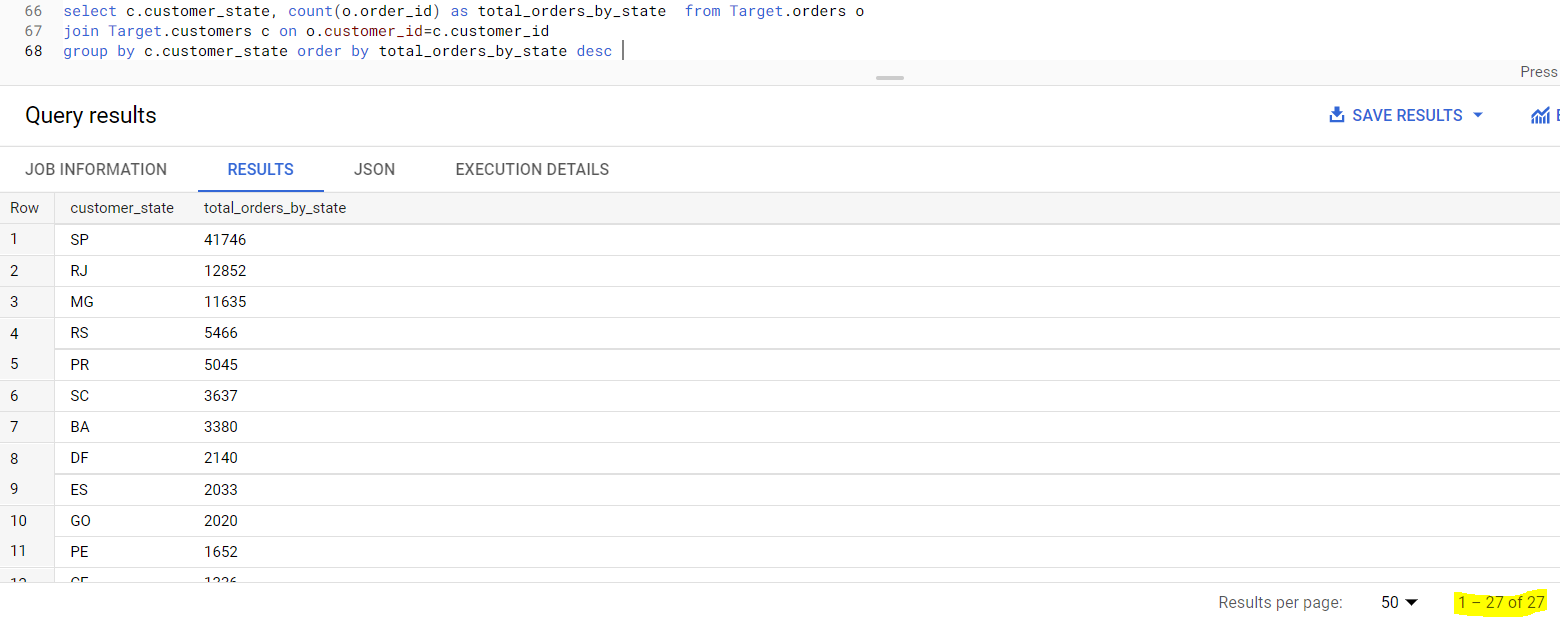
1. **Evolution of Ecommerce orders in Brazil region.**
2. Month on month orders by region:

We don’t know what region could be so we will do one city wise and another state wise.

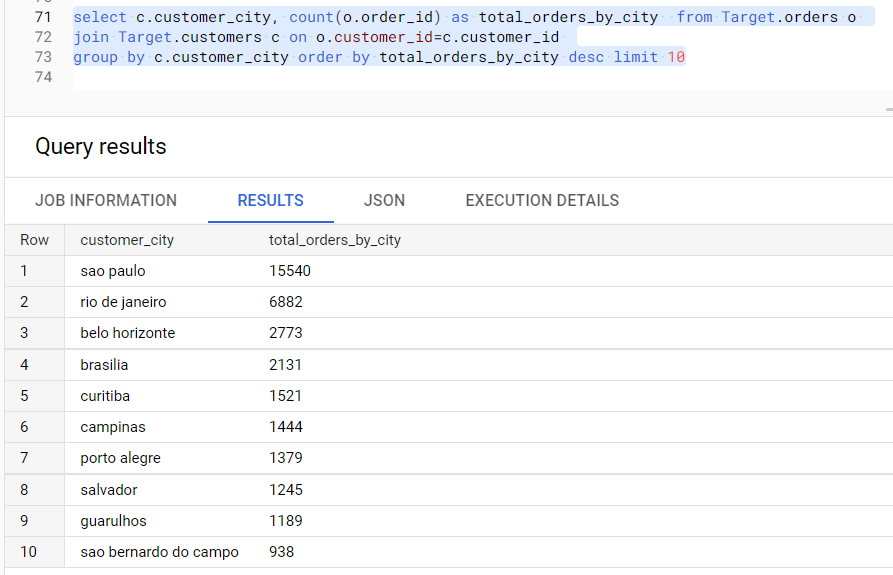
City wise orders: Sao paulo has the highest orders.



State wise orders: Out of 27 states, SP has the highest orders.

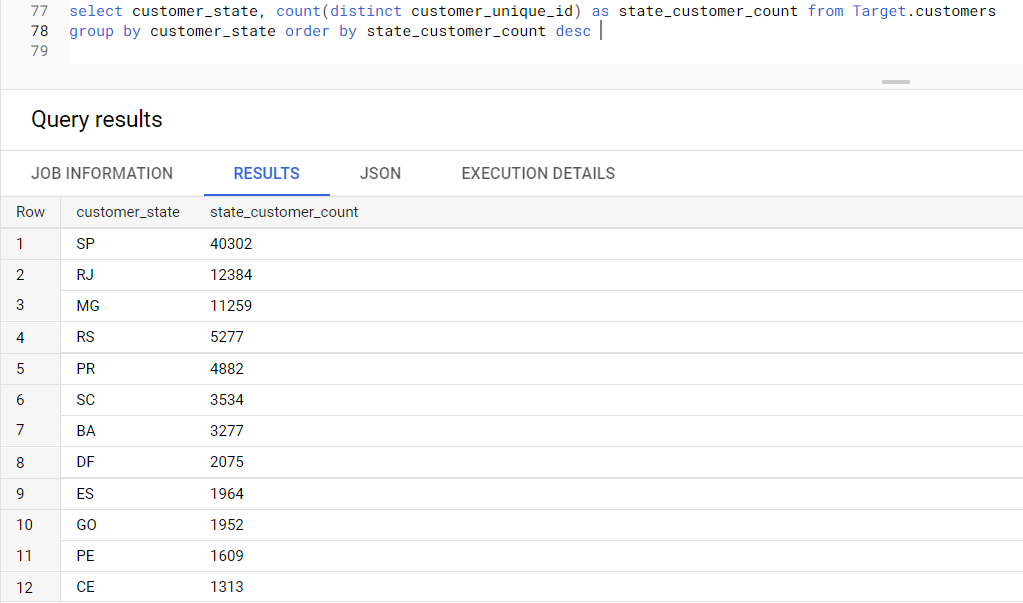


1. Top 10 cities with most orders.

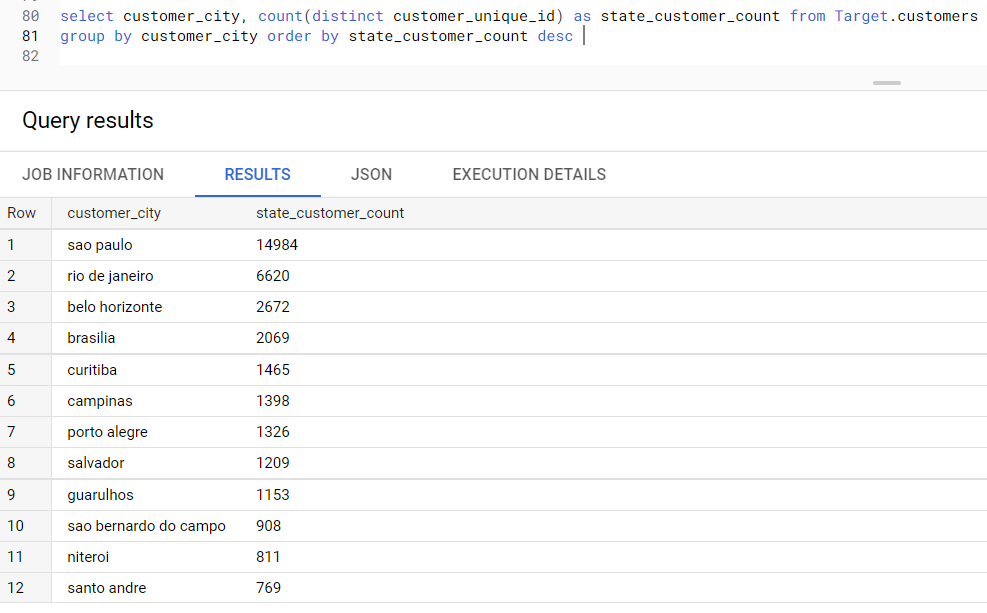


1. How are customers distributed in Brazil? State and city wise.

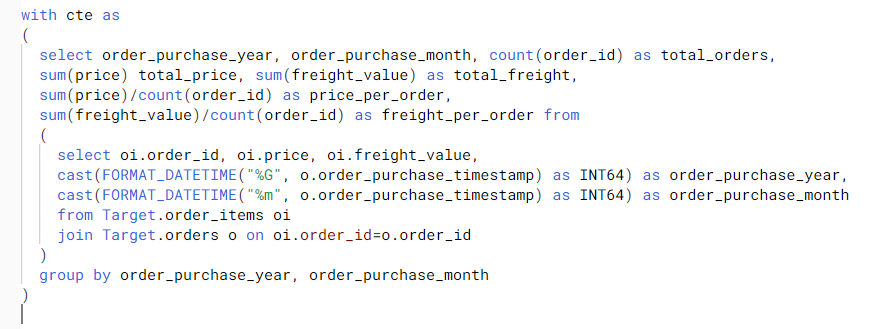
State wise: State SP has the highest number of customers. Here we have to take distinct customer\_unique\_id as it is duplicated in customers table.



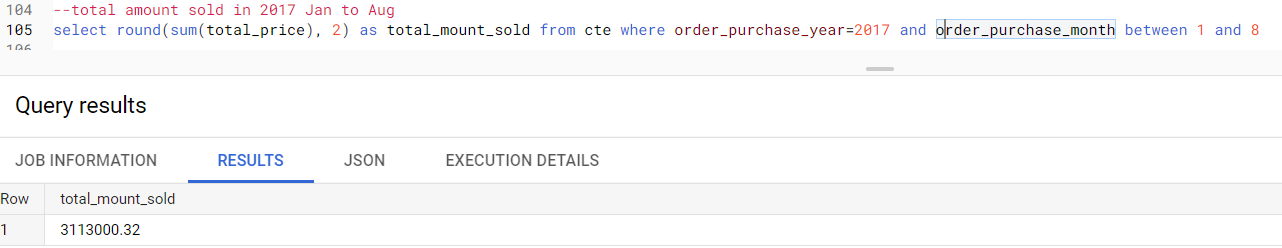
City wise: Sao paulo has the most number of unique customers.



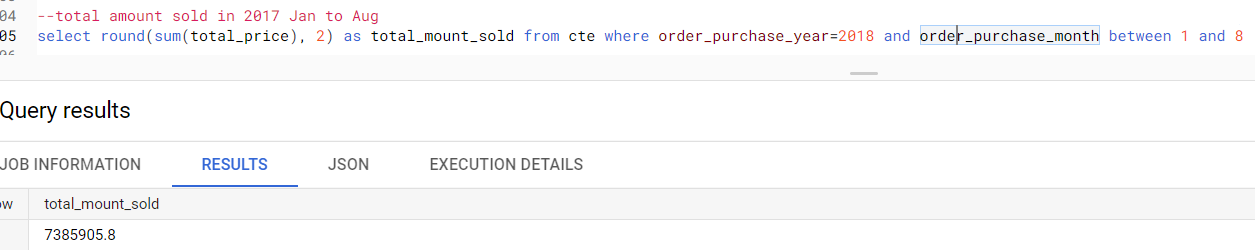
1. **Impact on Economy: Analyze the money movement by e-commerce by looking at order prices, freight and others.**
2. Created CTE:



1. Total amount sold in 2017 Jan to Aug

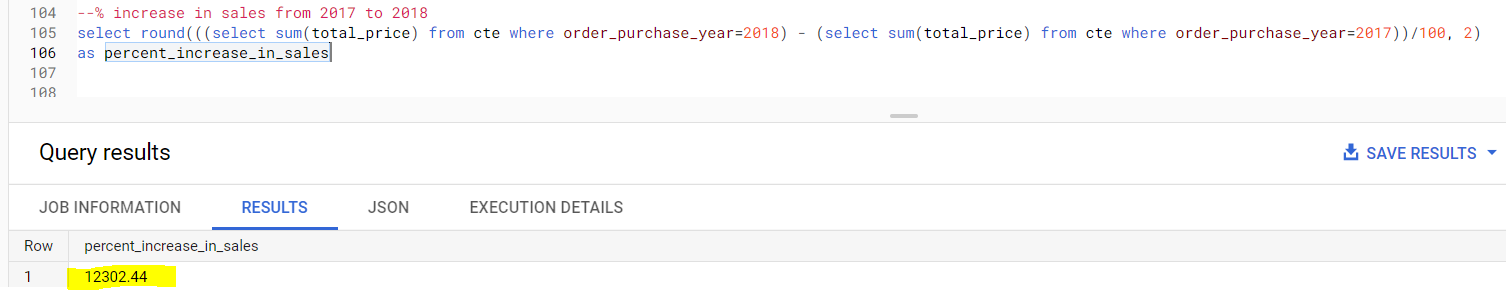


1. Total amount sold in 2018 Jan to Aug

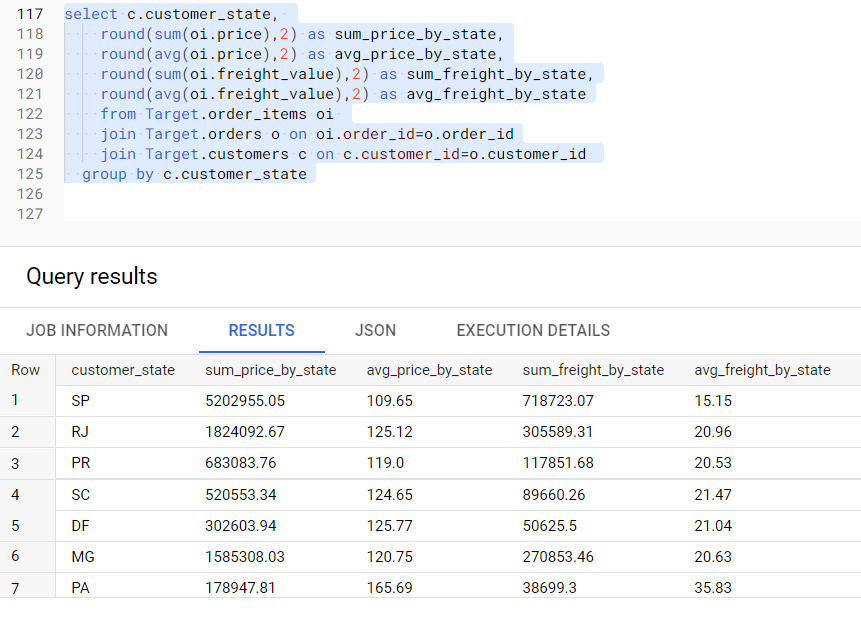


1. % increase in sales from 2017 to 2018

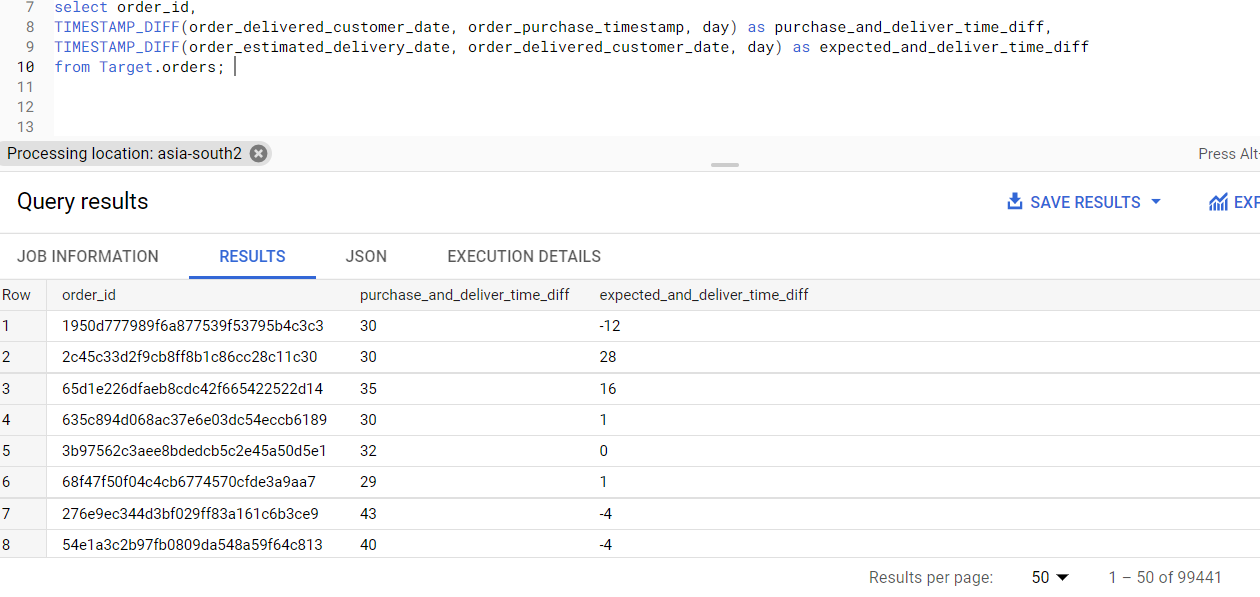
It is 12302 % increase, or 123.02 times increases in sales / revenue



1. Join with customer table and find Mean and Sum Prices & Freight by Customer state

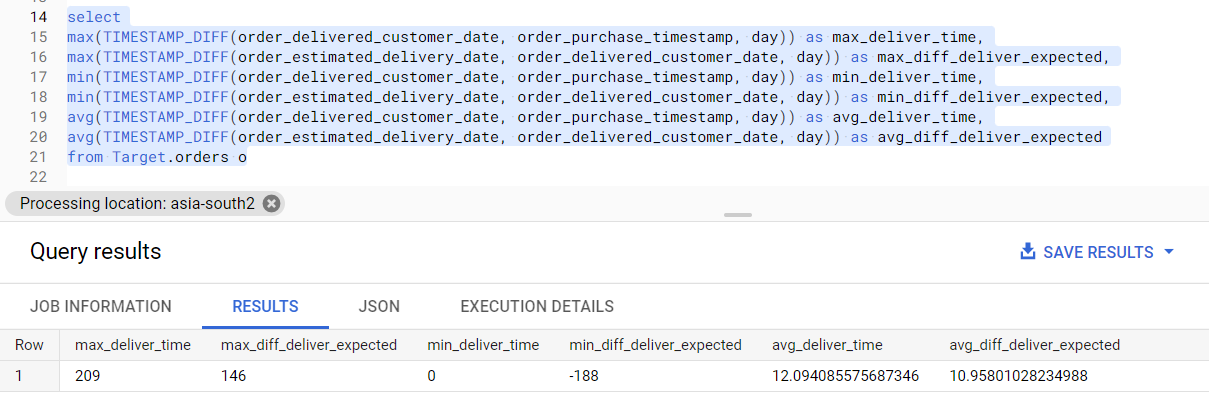


1. **Analysis on sales, freight and delivery time**
2. Calculating days between purchasing, delivering, estimated delivery

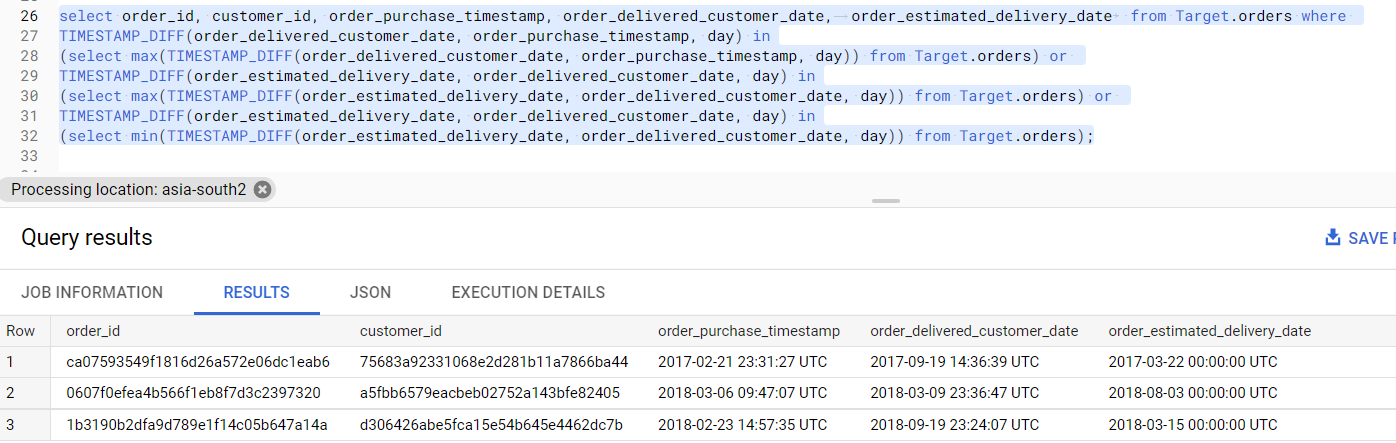


If we see we have some orders which took 209 days to deliver, some as small as 0.

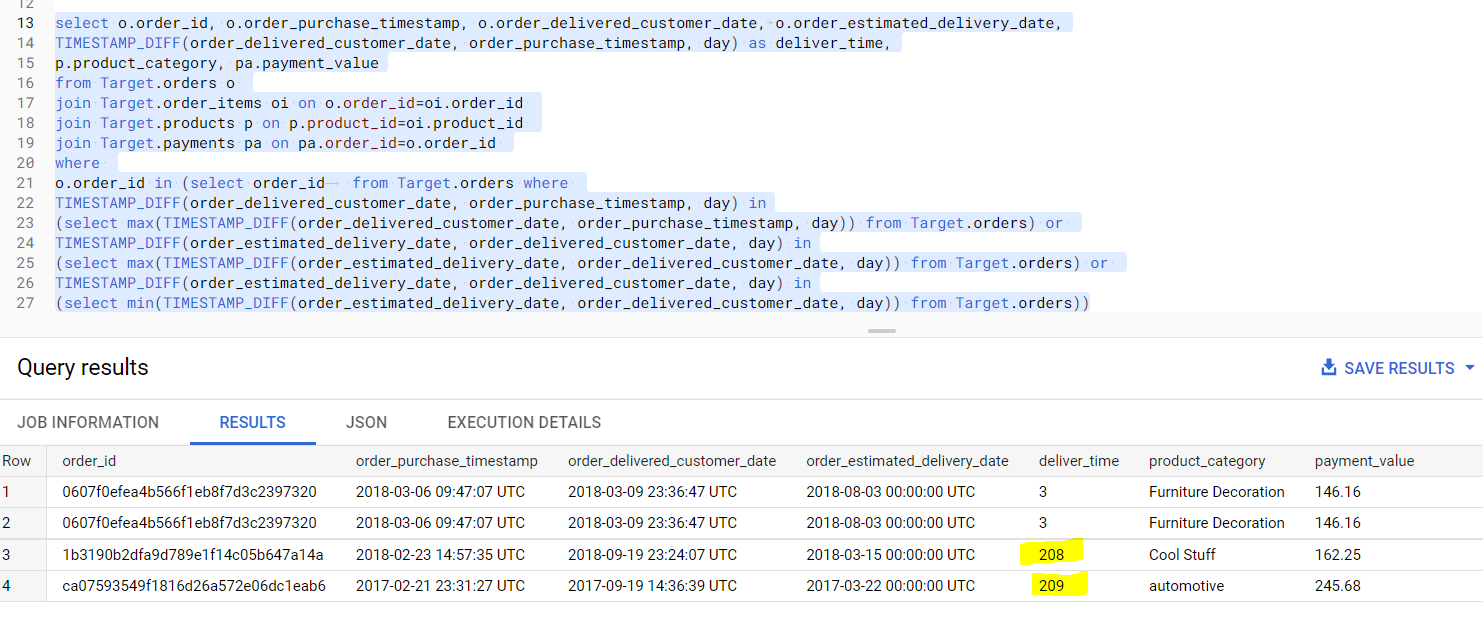
And max delivery time to expected time is 188 days(order was delivered after 188 days of expected date), min is 146(order was delivered 146 before the estimated date)



Let us analyze these 3 orders why such a vast time difference.

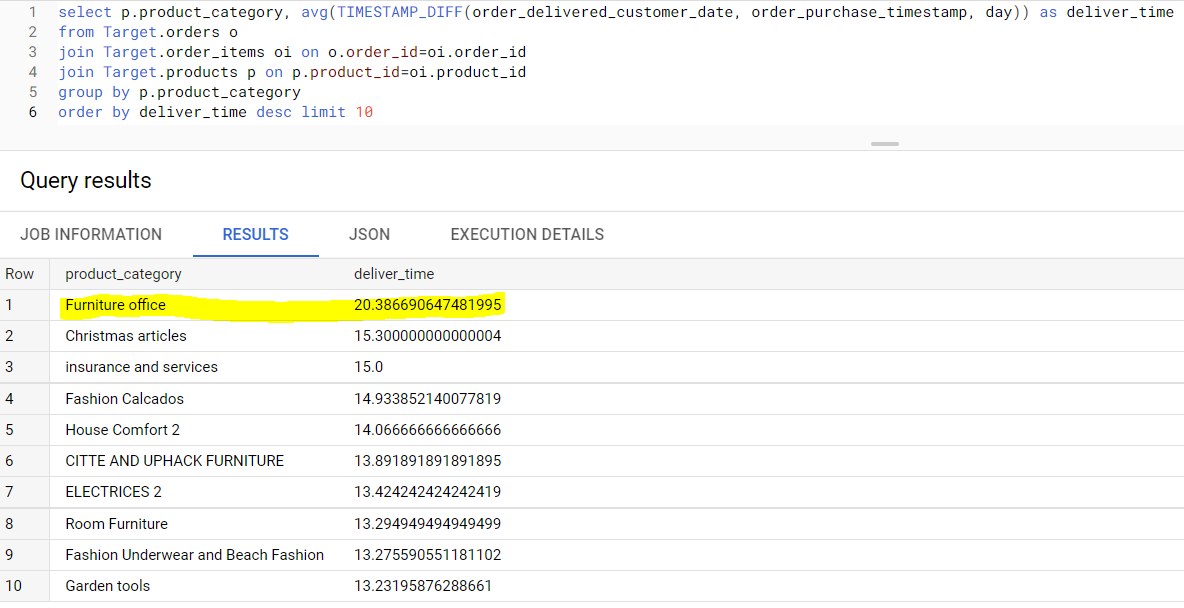


We see that for the first 2 rows are of same order id and they had very high expected date but was delivered quickly. Similarly for the 2nd and 3rd records the delivery time was more than 200 days. May be these are customized products (as on is automative and another is some cool stuff).



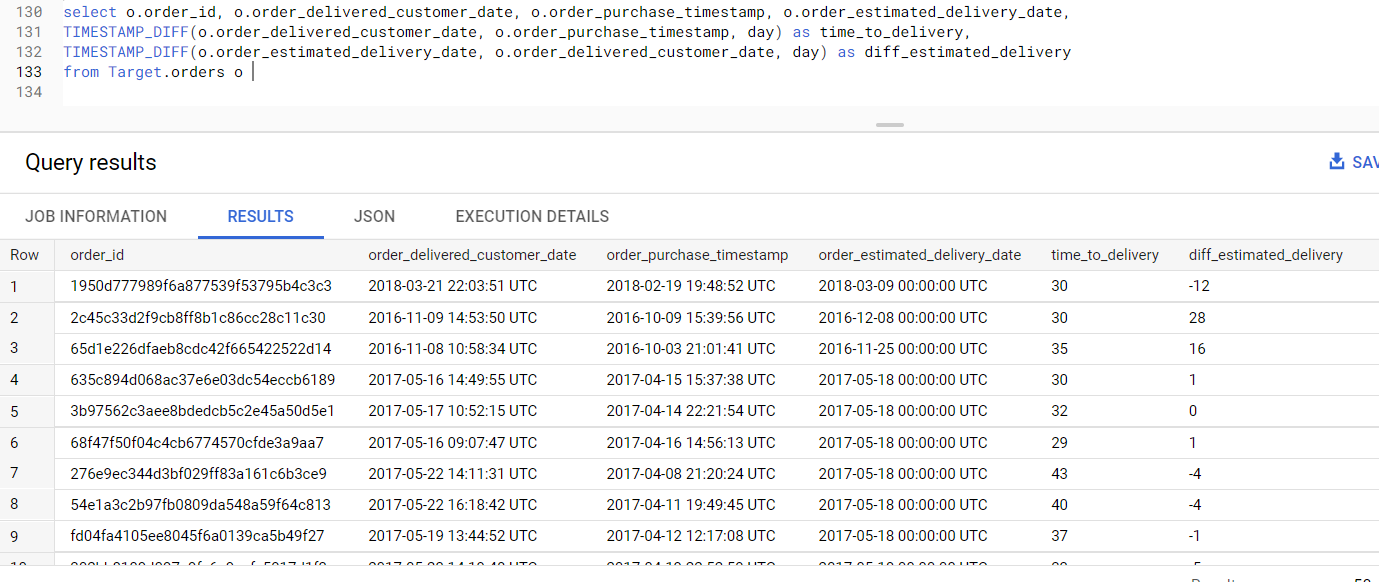
Let us check top 10 product categories for which delivery time is high.

We see that office furniture and Christmas articles, insurance and services are having high delivery time.



1. Create columns time\_to\_delivery, diff\_estimated\_delivery

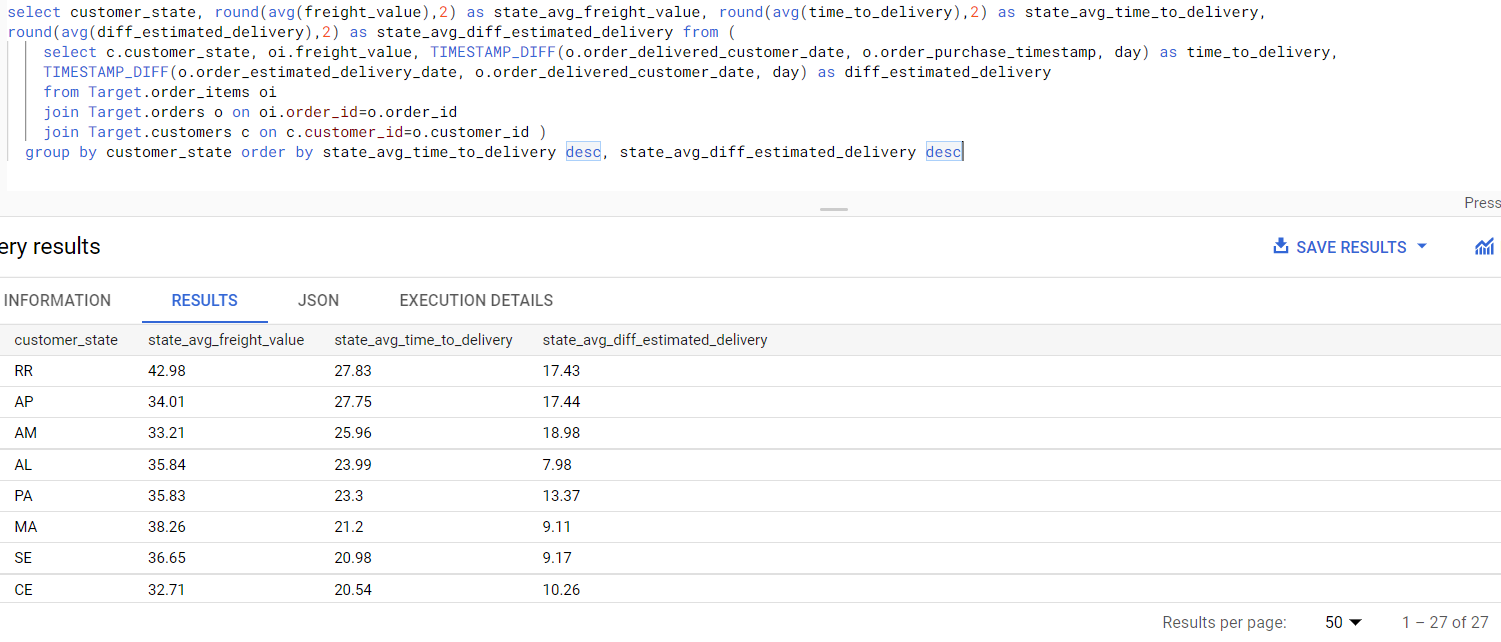
NOTE: Since I am using BigQuery free version I can not run ALTER table statements. So will just try to query these 2 columns instead of creating those new columns.



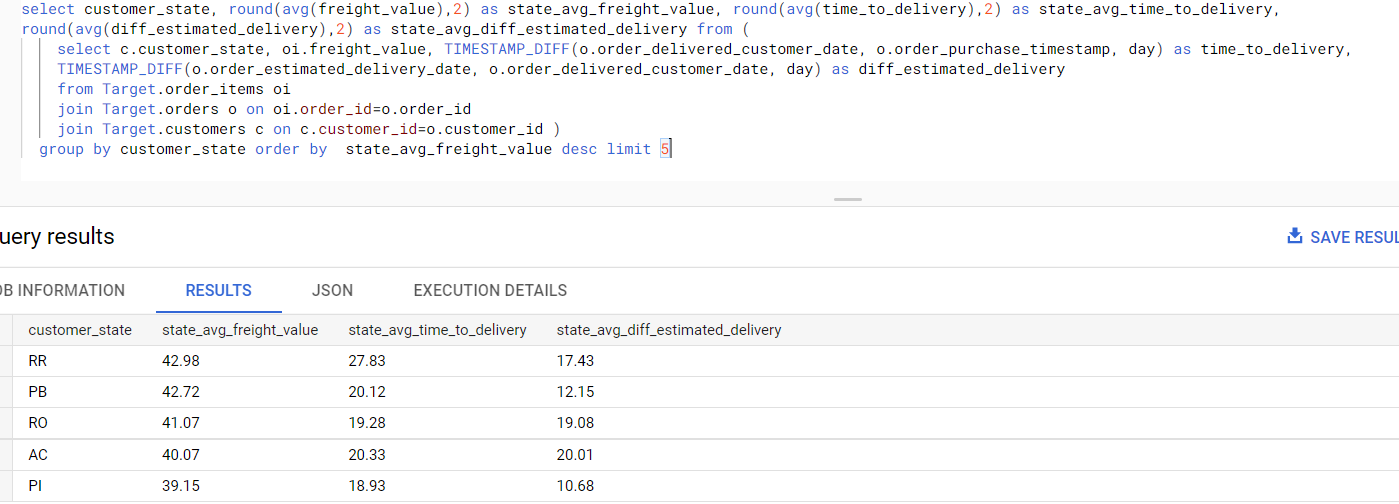
1. Grouping above analysis by state: State RR has the highest time of avg delivery time in days.

State SP has the least avg time to deliver orders.

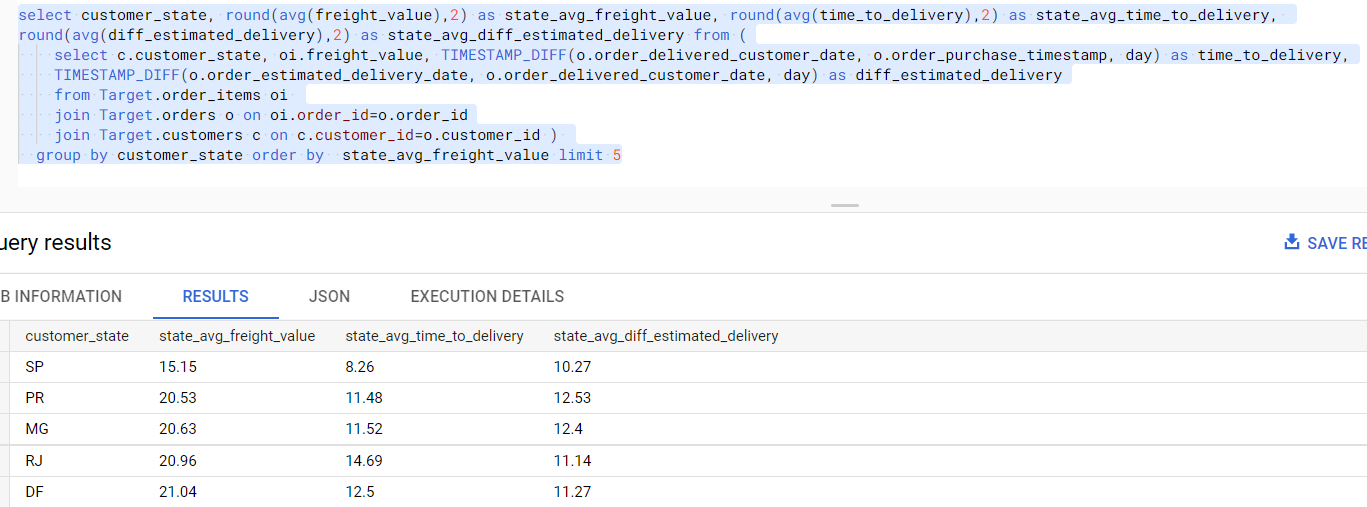
State AC tops in time difference to delivery and estimated time to delivery, State AL lies at the bottom in this category.



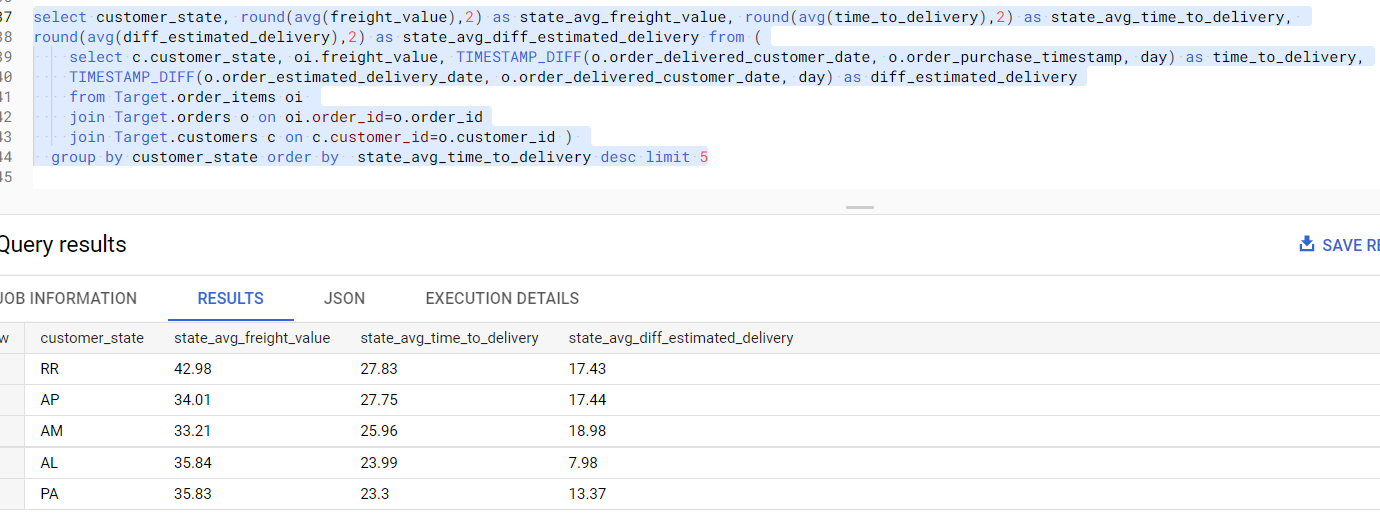
1. Top 5 states with highest avg freight value



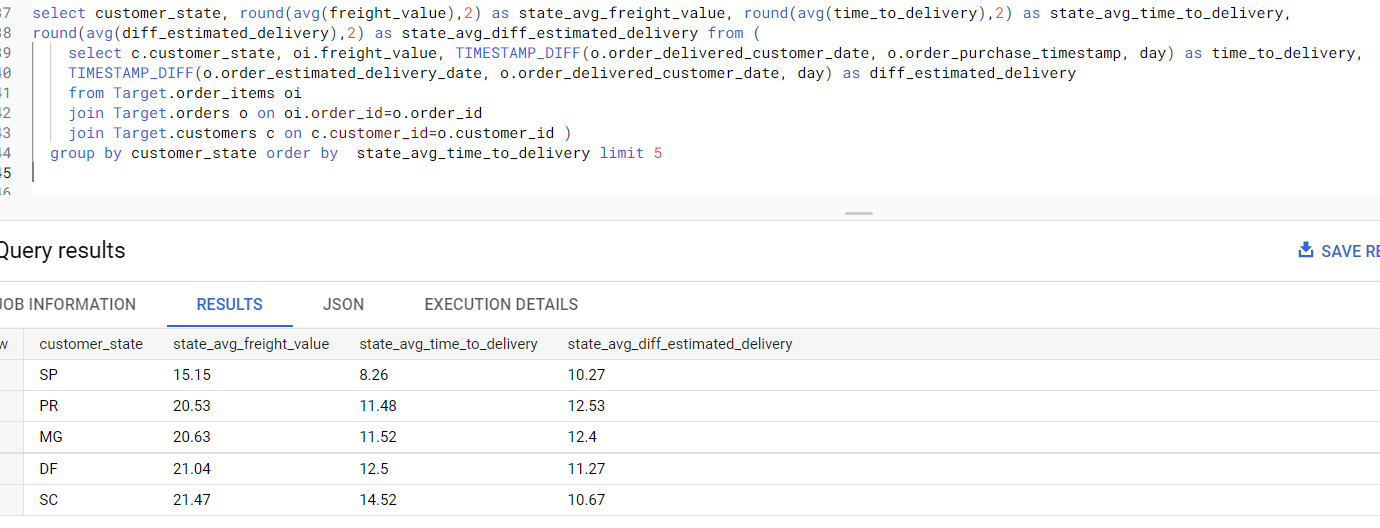
Top 5 states with lowest avg freight value



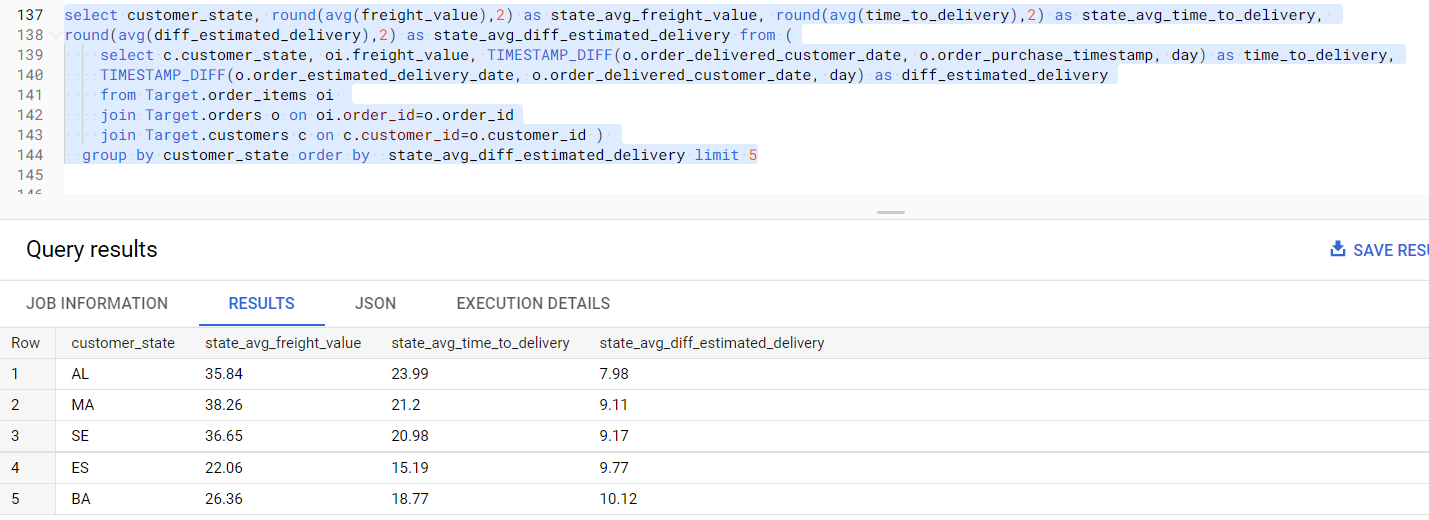
1. Top 5 states with highest avg time to delivery



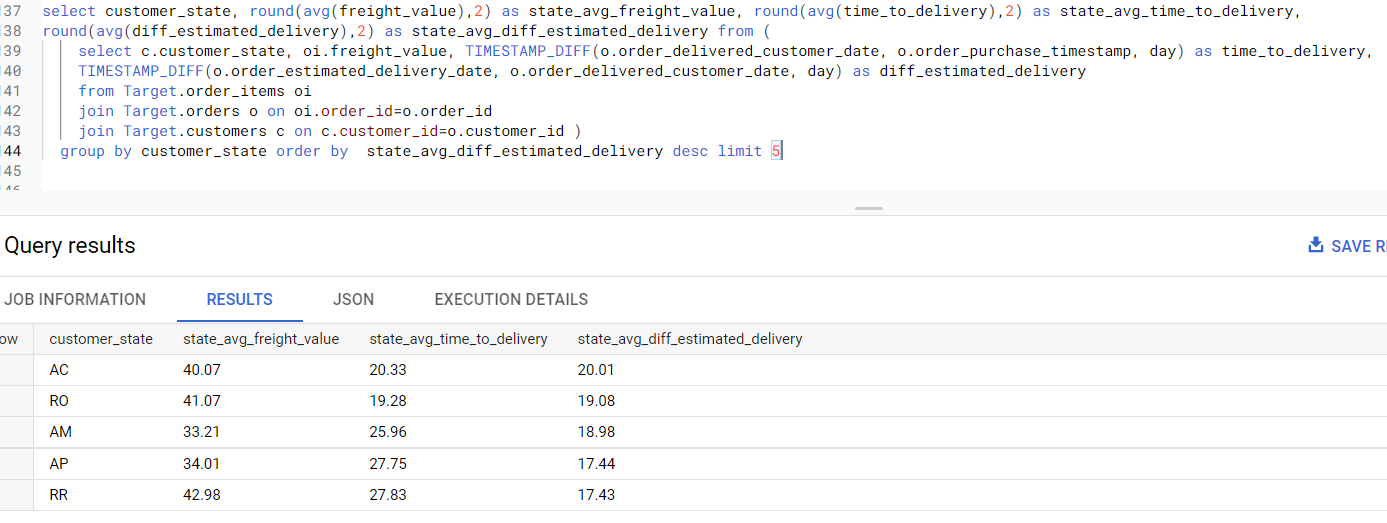
Top 5 states with lowest avg time to delivery



1. Top 5 states with delivery fast as compared to estimated date



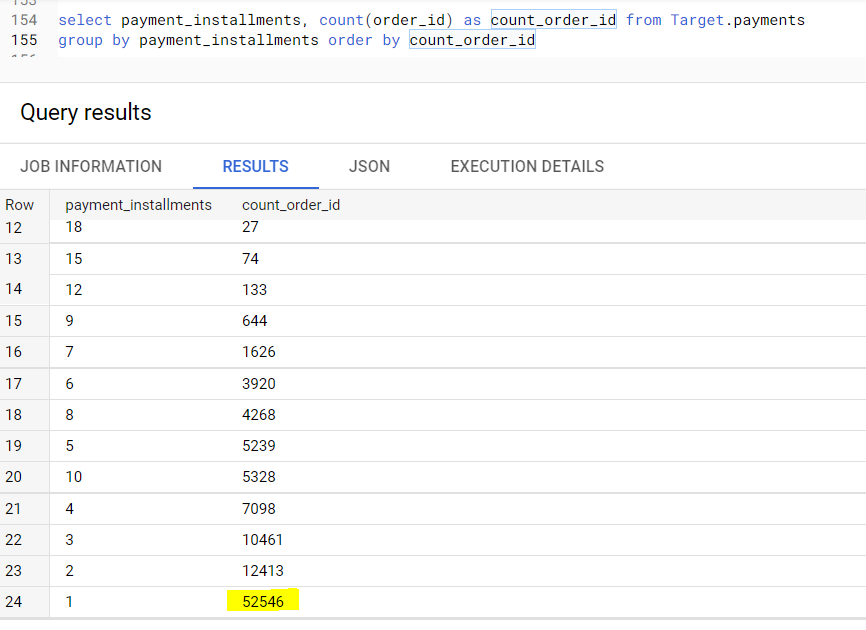
Top 5 states with delivery slow as compared to estimated date



1. **Payment type analysis: Join “payments” dataset with the existing data on order\_id**
2. Count of orders for different payment types : Most people prefer to pay using credit cards, some using UPI.

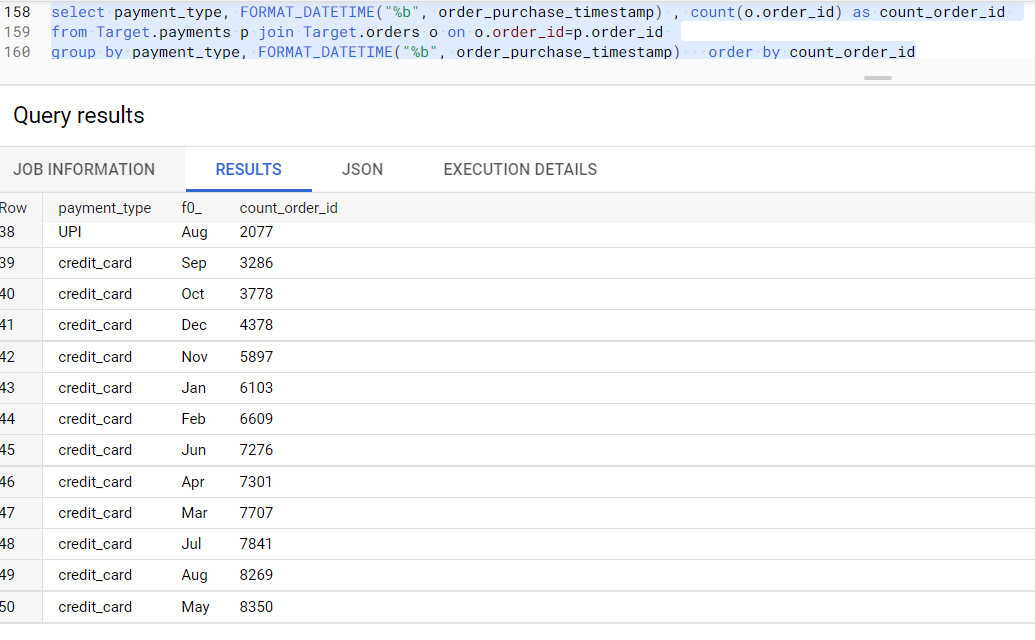


1. Distribution of payment\_installments and count of orders. Almost 50 % of people pay in just one go, they don’t opt for EMI.

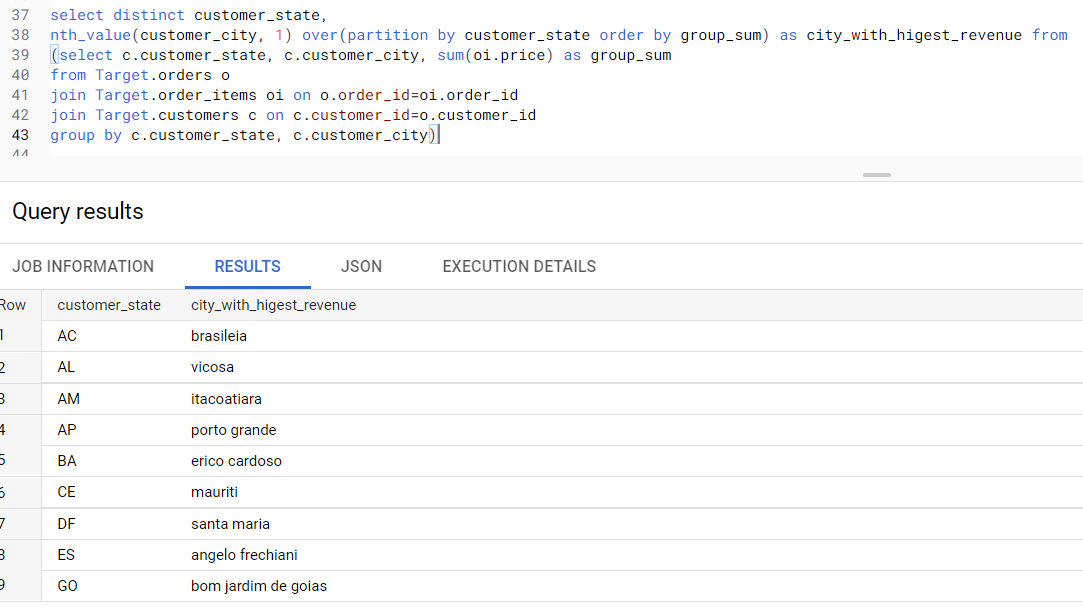


1. Count of order for different payment types month over month

Credit card usage is high in the moth of May. UPI has the highest usage in August.



1. Extra Analysis:
2. Find top cities in each states based on their revenue.



1. Show the total payments received cumulative over each year and month / the total revenue till date for each year and month



1. **Insights:**
2. Number of orders keeps on decreasing as the week progresses. Weekends has very less orders compared to Monday and Tuesday.
3. Number of orders are high during the time period 12 Noon to 6PM.
4. Though the number of orders decreased from 2018 Q1, still the overall price of products sold / revenue still remains high as compared to 2017. This shows a lot of orders are for multiple products on 2018. Which is a good sign.
5. States RR, PB, RO, AC, PI has the highest freight price and SC, PR, MG, RJ, DF has the lowest freight price.
6. RR, AP, AM, AL, PA states has higher avg delivery time. SC, DF, MG, PR, SP has the lowest delivery times. This difference is so bad that the avg delivery time difference between RR and SP is 20 days.
7. AL, MA, SE, ES, BA has the lowest and AC, RO, AM, AP, RR has the highest estimated to delivery time difference.
8. **Actions on Business:**
9. We saw that the unique customers were increasing quarter wise from 2016 Q4 till 2018 Q1, then it got decreased for 2018 Q2 and Q3. Business needs check what causes this issue in acquiring customers.
10. As weekends are having less orders we need to have some super sales day on weekends which will make customers buy products in weekends.
11. We saw that as the number of customers from a state increases the orders also increases, so business needs to focus on sates where customer participation is less. Few such states are RR, AP, AC.
12. States RR, PB, RO has the highest freight price, We need to figure out ways to lessen freight by finding sellers nearer to these states / by improving supply chain for these states.
13. RR, AP, AM, AL these states has higher expected to delivery time difference, this could be one of the reasons of these states also having higher freight value.
14. It is good to see that almost 75% of the orders are being paid through credit cards. This can also help us if we partner with most used credit cards banks and provide some discount on usage of those cards.
15. We see that office furniture and Christmas articles, insurance & services are having high delivery time. We need to identify vendors for these product categories across different demographic regions, which would help in faster shipping, hence less delivery time.