HIVE Queries

1. **What are the top 10 stations where bikes are picked from (starting station for a trip)?**

**Query:**

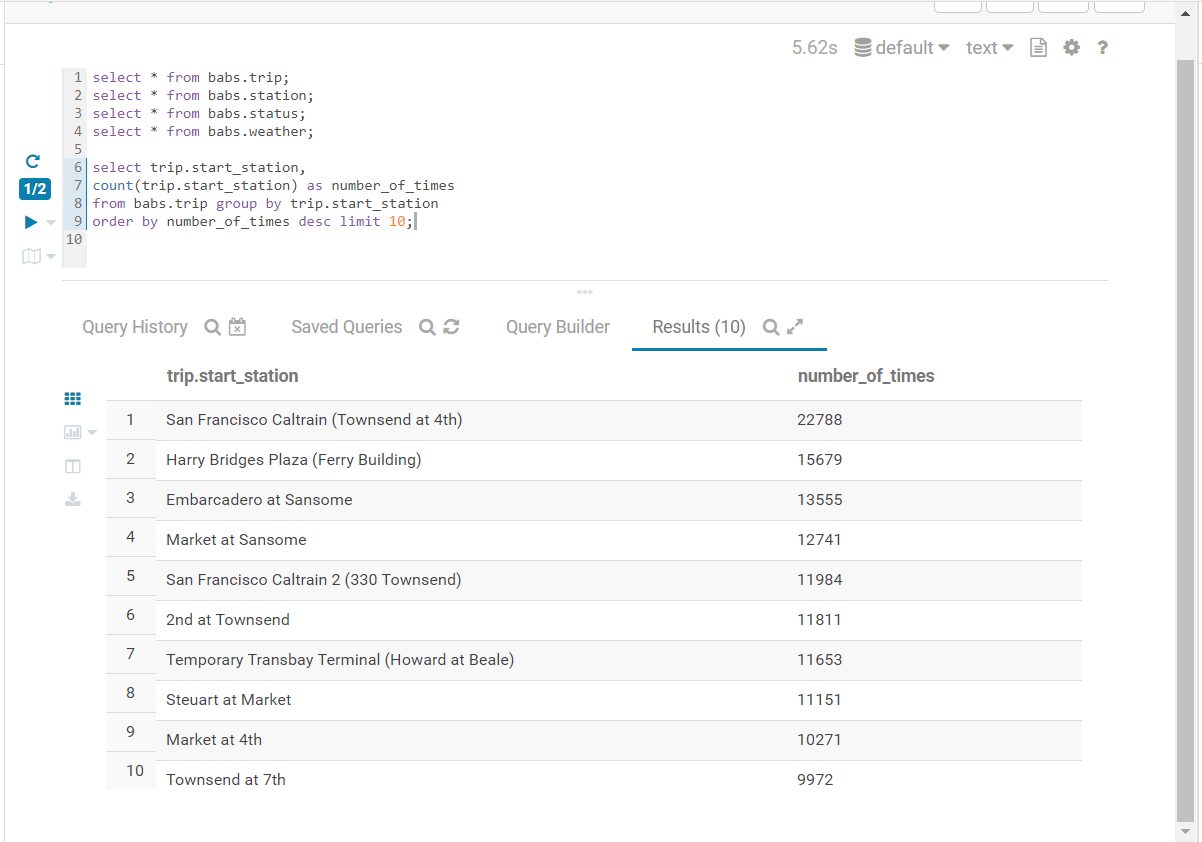
select trip.start\_station,

count(trip.start\_station) as number\_of\_times

from babs.trip group by trip.start\_station

order by number\_of\_times desc limit 10;

**Output:**



1. **What are the 10 most popular trips (pairs of starting and end locations)?**

**Query:**

select trip.start\_station, trip.end\_station,

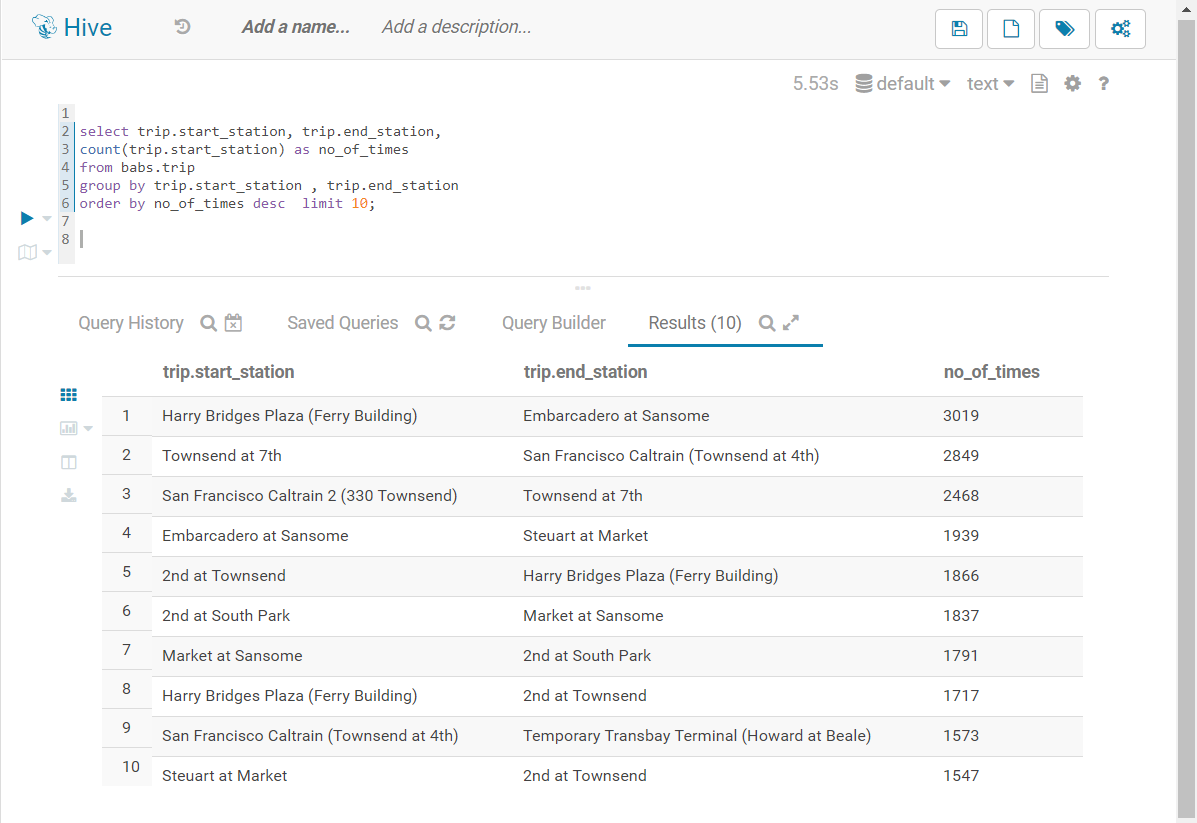
count(trip.start\_station) as no\_of\_times

from babs.trip

group by trip.start\_station , trip.end\_station

order by no\_of\_times desc limit 10;

**Output:**



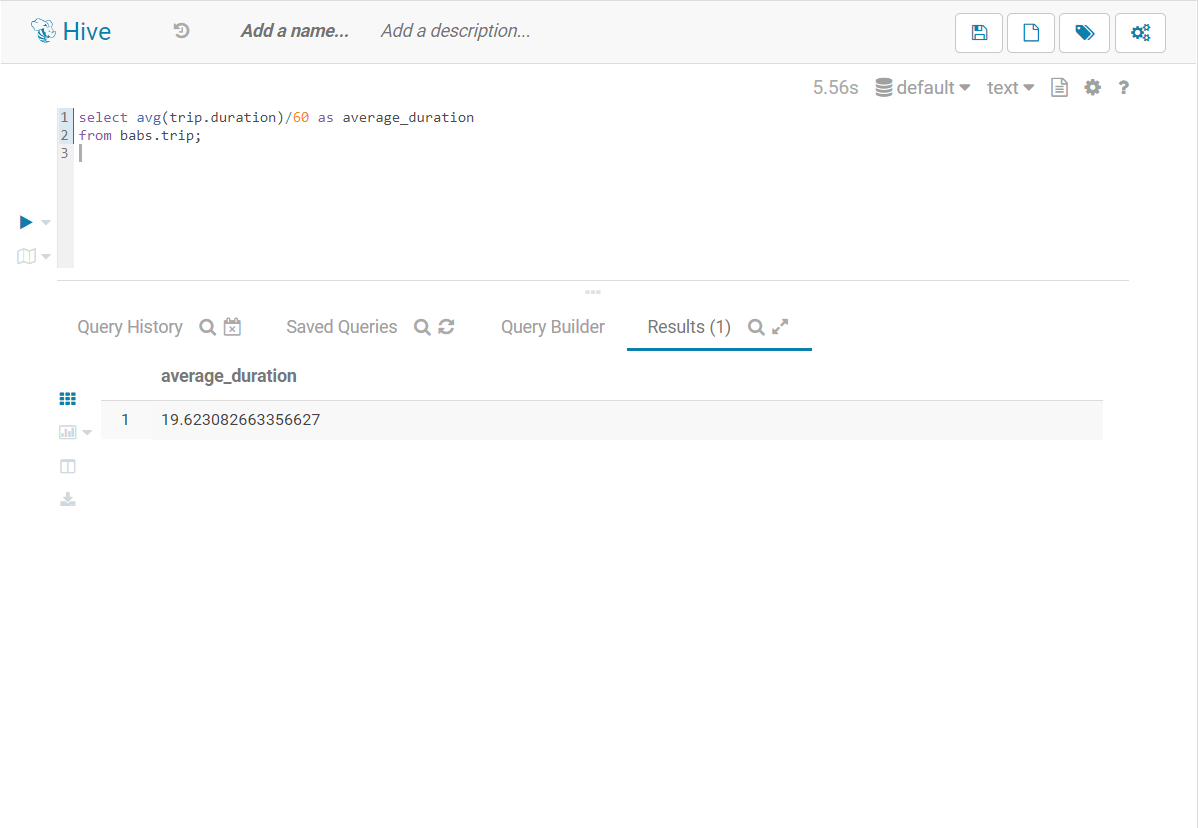
1. **What is the average duration of a trip in minutes?**

**Query:**

select avg(trip.duration)/60 as average\_duration

from babs.trip;

**Output:**



1. **Which are the 10 bikes which have been used the most?**

**Query:**

select trip.bike\_id,

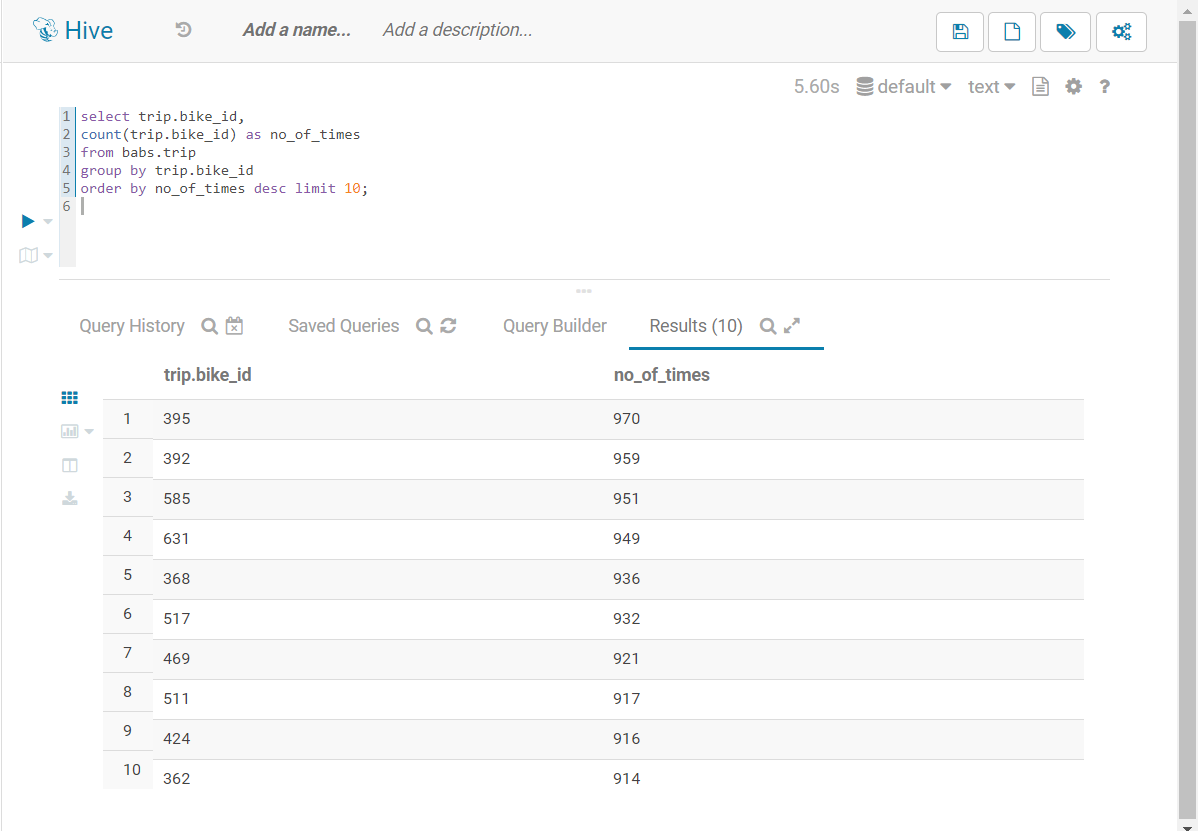
count(trip.bike\_id) as no\_of\_times

from babs.trip

group by trip.bike\_id

order by no\_of\_times desc limit 10;

**Output:**



1. **What are 10 most popular destinations from the station San Francisco Caltrain (Townsend at 4th)?**

**Query:**

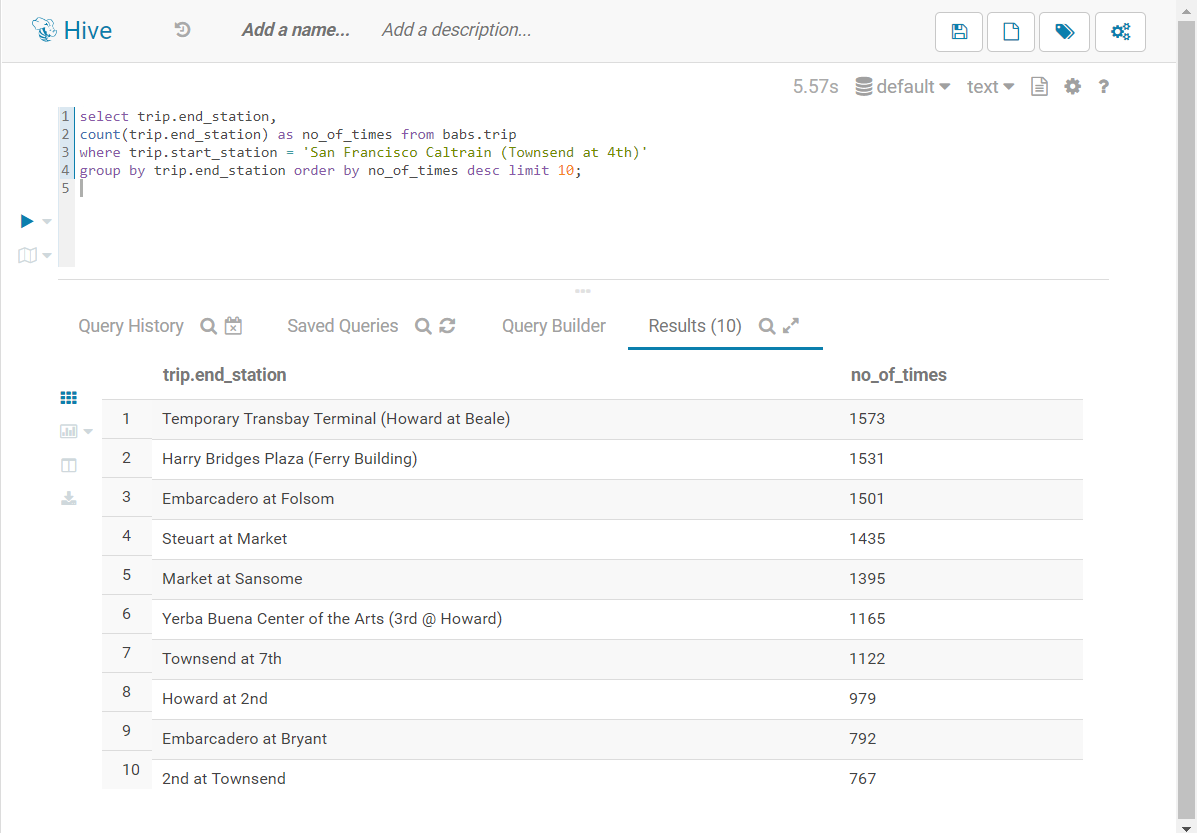
select trip.end\_station,

count(trip.end\_station) as no\_of\_times from babs.trip

where trip.start\_station = 'San Francisco Caltrain (Townsend at 4th)'

group by trip.end\_station order by no\_of\_times desc limit 10;

**Output:**



1. **Visualize the result of 5th question using Hue’s geospatial visualizations.**

**Query:**

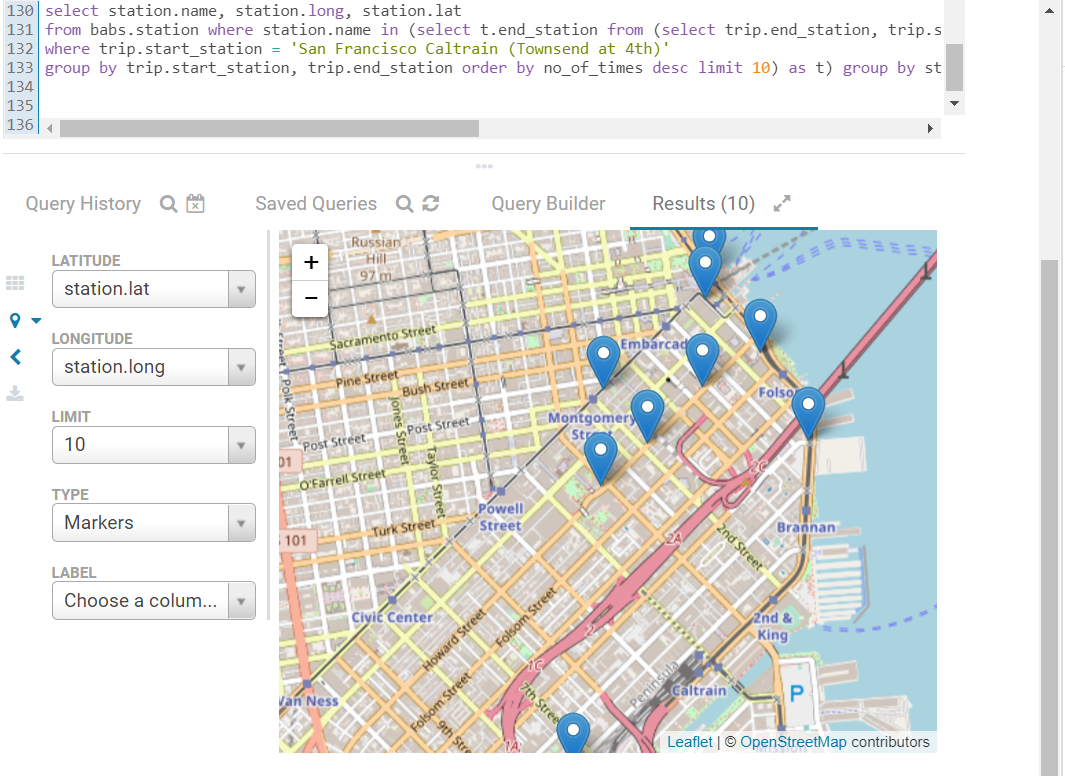
select station.name, station.long, station.lat

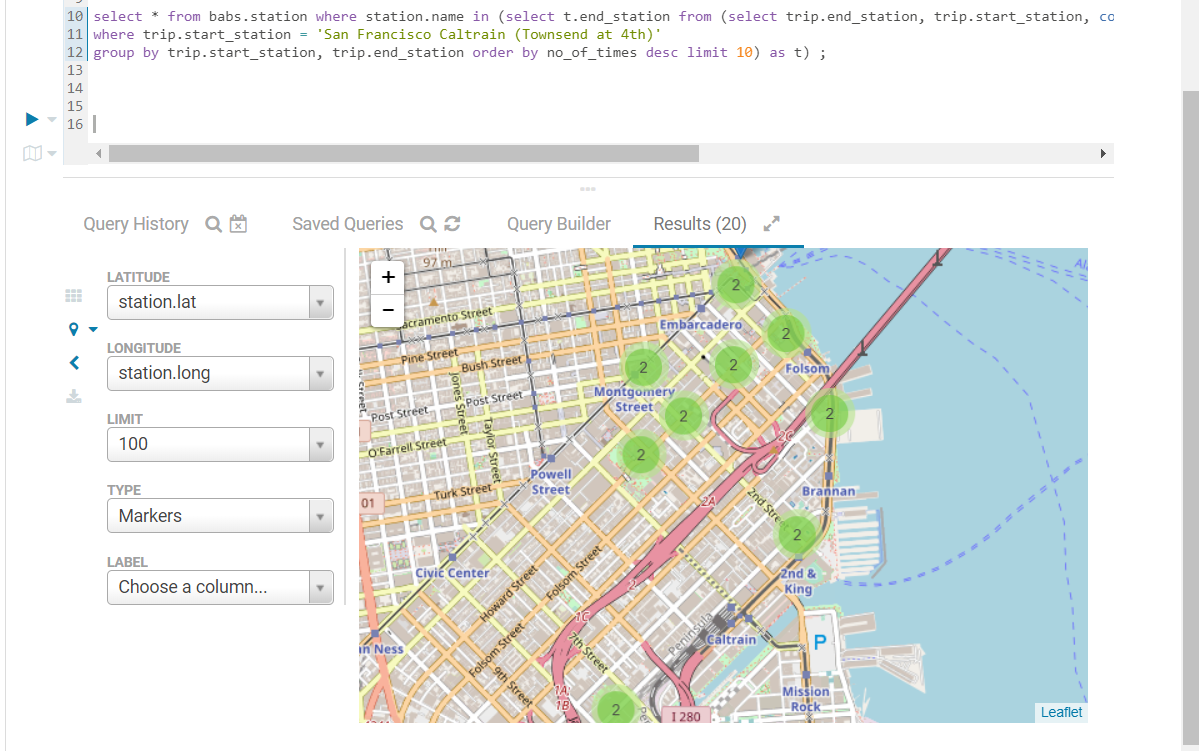
from babs.station where station.name in (select t.end\_station from (select trip.end\_station, trip.start\_station, count(\*) as no\_of\_times from babs.trip

where trip.start\_station = 'San Francisco Caltrain (Townsend at 4th)'

group by trip.start\_station, trip.end\_station order by no\_of\_times desc limit 10) as t) group by station.name, station.long, station.lat

**Output:**





1. **Visualize the total number of trips and average duration in minutes, grouped by hour for San Francisco Caltrain (Townsend at 4th) station using a scatterplot.**

**Query:**

select t.start\_hours as start\_hours ,

count(t.start\_hours) as total\_trips,

avg(t.trip\_duration)/60 as average\_duration,

round(avg(t.trip\_duration)/60) as average\_duration\_in\_round from

(select trip.duration as trip\_duration ,

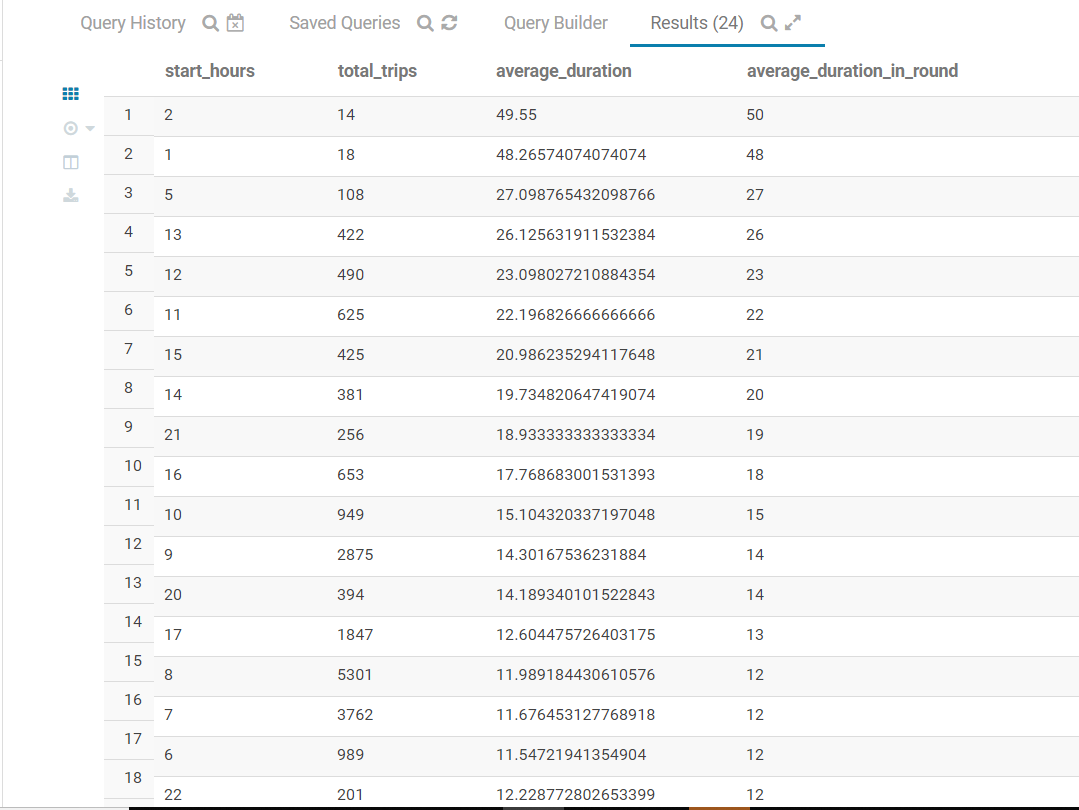
(from\_unixtime(unix\_timestamp(trip.start\_time , 'MM/dd/yyyy'), 'yyyy-dd-MM')) as start\_date,

hour((from\_unixtime(unix\_timestamp(trip.start\_time , 'MM/dd/yyyy HH:MM'), 'yyyy-dd-MM HH:MM:SS'))) as start\_hours from babs.trip

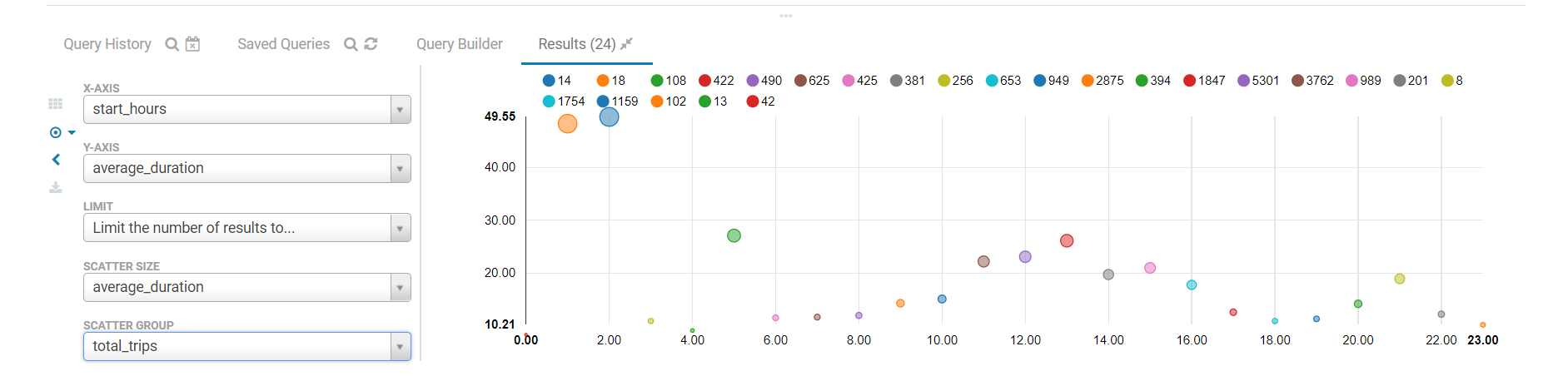
where trip.start\_station='San Francisco Caltrain (Townsend at 4th)') as t

group by t.start\_hours order by average\_duration\_in\_round desc;

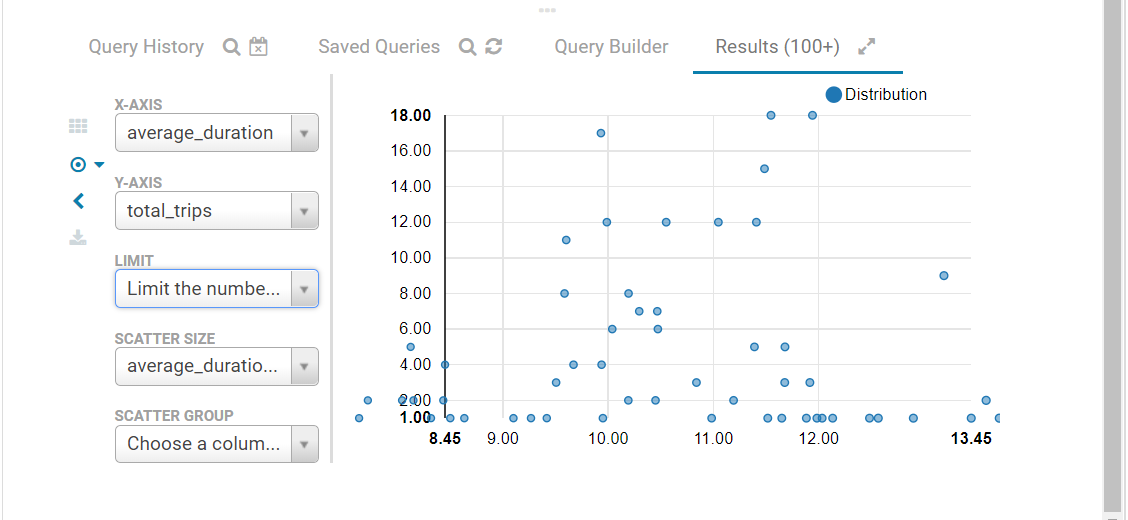
**Output:**

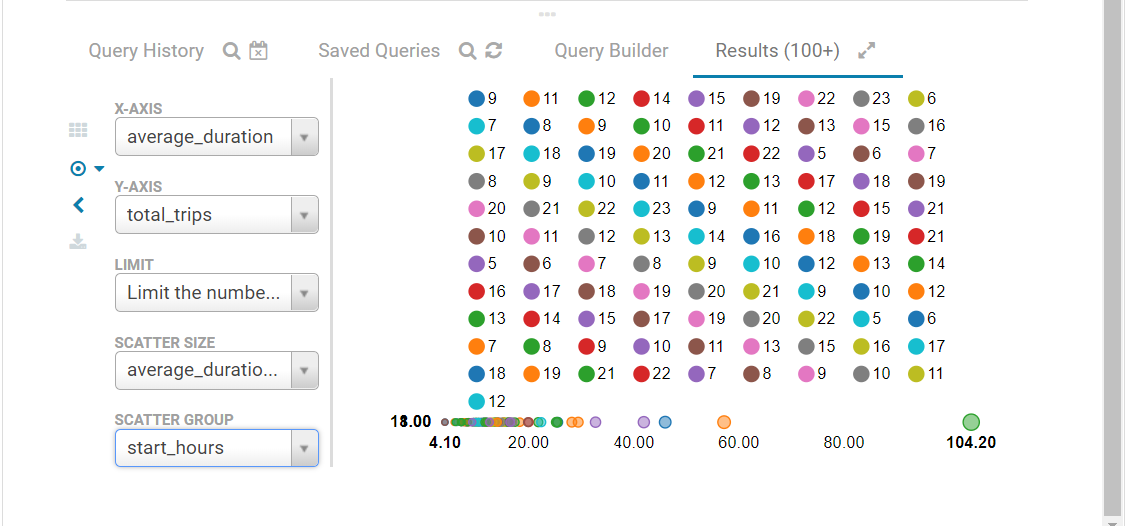


**Visualization:**



**The below visualization is with respect to start date**





1. **How many trips were charged with overtime fees? (Trips having duration over 30 minutes)**

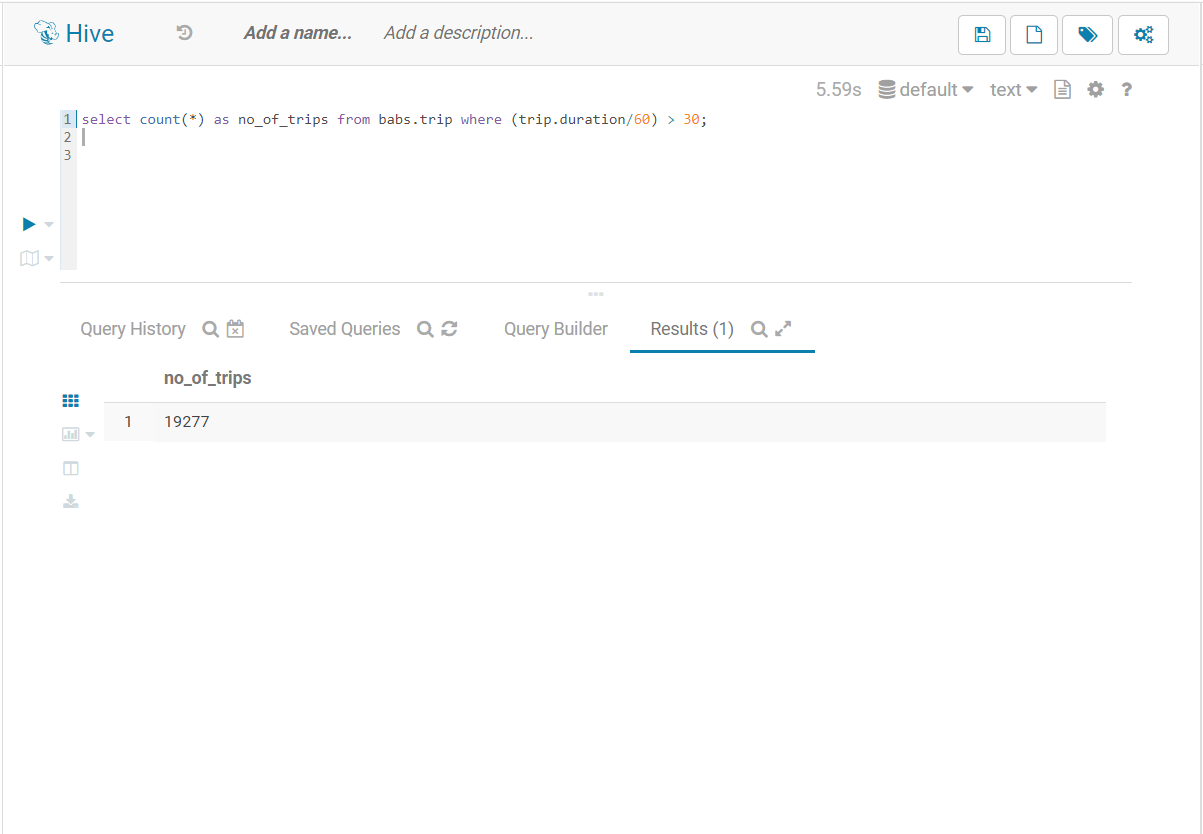
**Query:**

select count(\*) as no\_of\_trips

from babs.trip

where (trip.duration/60) > 30;

**Output:**



1. **Which trips cause highest number of overtime charges, how many seconds in total overtime was charged?**

**Query:**

select trip.start\_station as start\_station, trip.end\_station as end\_station,

count(\*) as number\_of\_trips, round(avg(trip.duration)) as average\_duration,

round(avg(trip.duration)-1800) as avg\_overtime\_seconds,

(sum(trip.duration) - (1800 \* count(\*))) as total\_overtime\_sec,

(round(avg(trip.duration)-1800) \* count(\*)) as toal\_OT\_Sec\_usingAvg

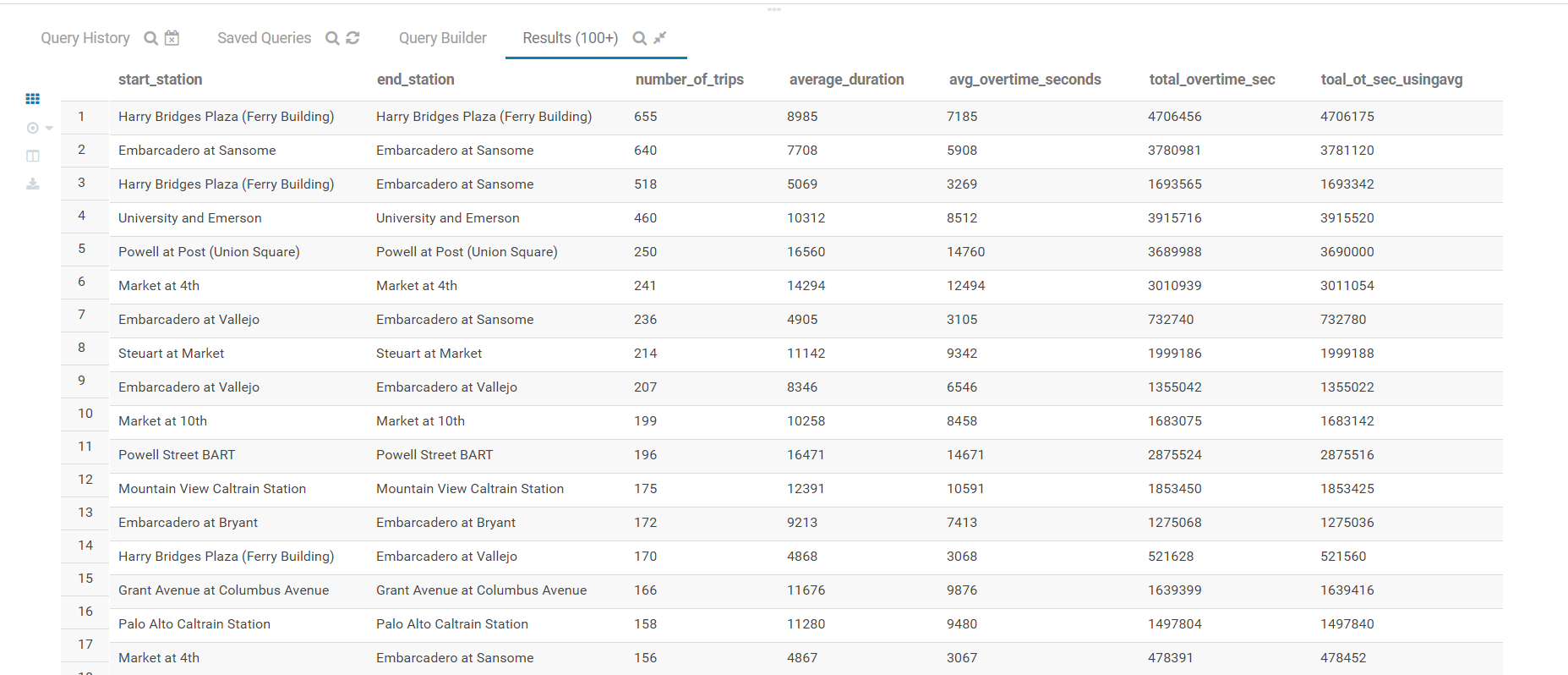
from babs.trip

where trip.duration > 1800

group by trip.start\_station, trip.end\_station

order by number\_of\_trips desc;

**Output:**



1. **Visualize the breakdown of trips by subscription type.**

**Method 1:**

**Query:**

select t.subscription\_type, start\_day, count(\*) as TripCount from

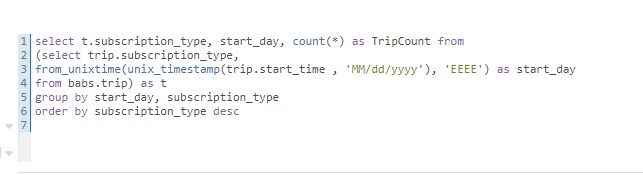
(select trip.subscription\_type,

from\_unixtime(unix\_timestamp(trip.start\_time , 'MM/dd/yyyy'), 'EEEE') as start\_day

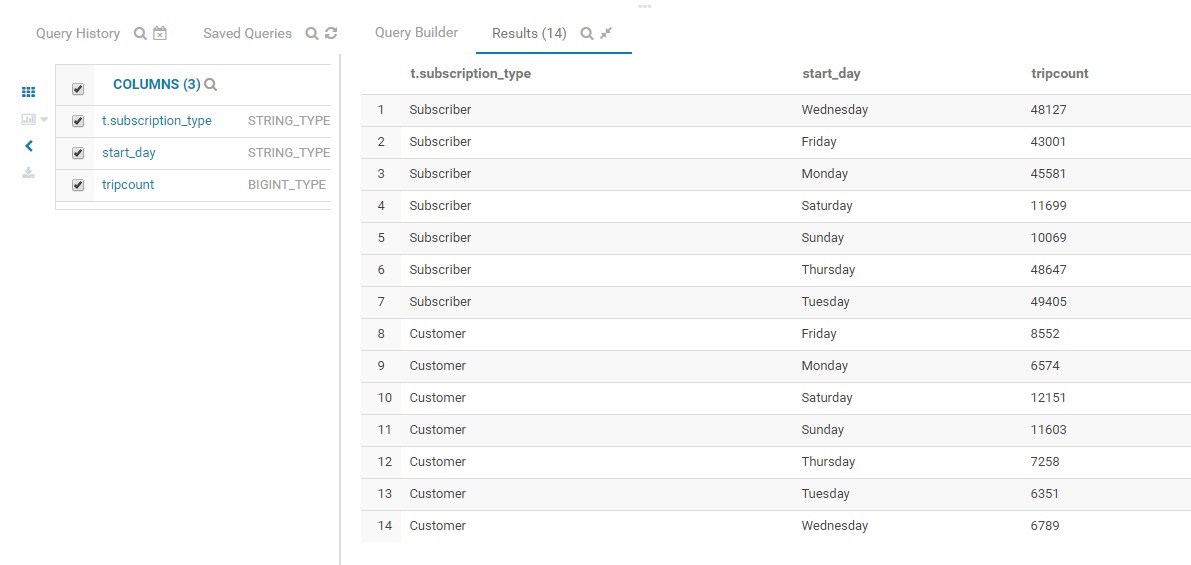
from babs.trip) as t

group by start\_day, subscription\_type

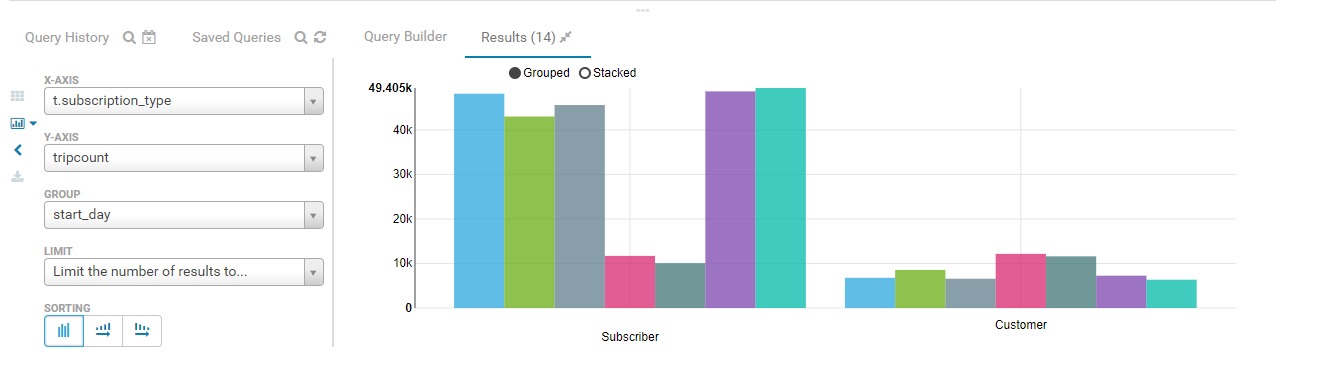
order by subscription\_type desc



**Output:**



**Visualizations:**



**Method 2:**

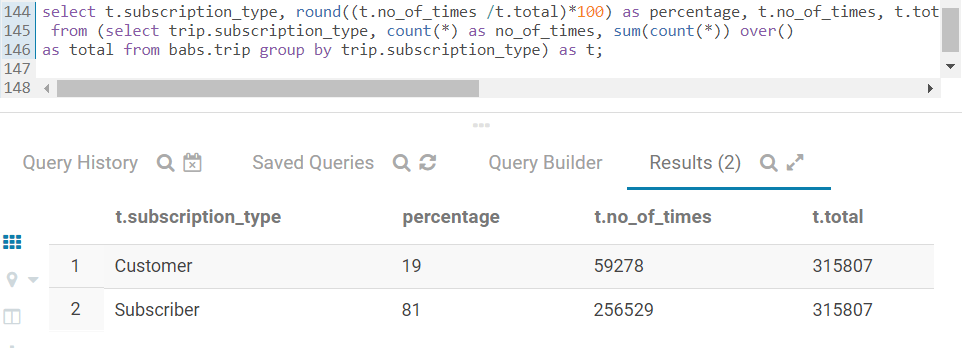
**Query:**

select t.subscription\_type, round((t.no\_of\_times /t.total)\*100) as percentage, t.no\_of\_times, t.total

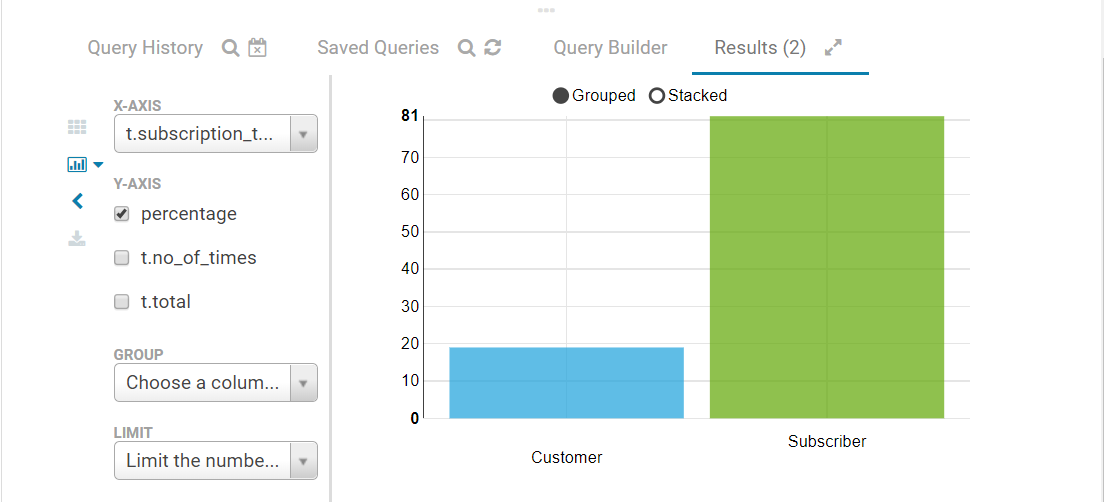
from (select trip.subscription\_type, count(\*) as no\_of\_times, sum(count(\*)) over()

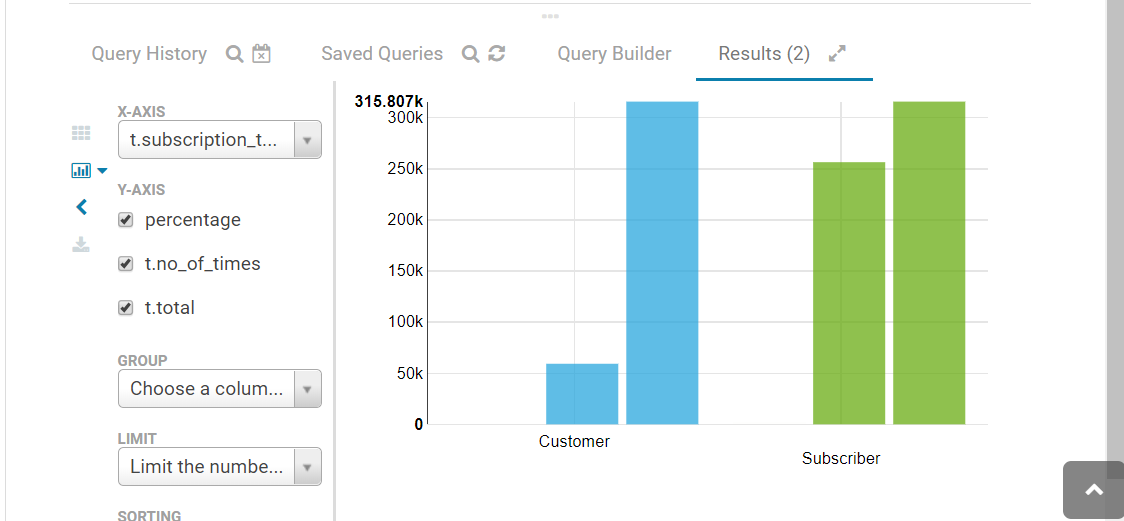
as total from babs.trip group by trip.subscription\_type) as t;

**Output:**



**Visualizations:**





1. **The Bike Share system is intended to be used for short rides: trips under a half hour do not incur any additional charges. Do riders use the system in the intended way? (Do subscribers make many shorter trips compared to customers who do not pay monthly subscription?)**

**Query:**

create table tempTable(

subscription\_type varchar(30),

condition\_type varchar(30)) ;

insert into tempTable

(

select trip.subscription\_type,

case

when (trip.duration/60) > 30 then 'greaterthan30'

when (trip.duration/60) < 30 then 'lessthan30'

when (trip.duration/60) = 30 then 'equals30'

end as condition\_type

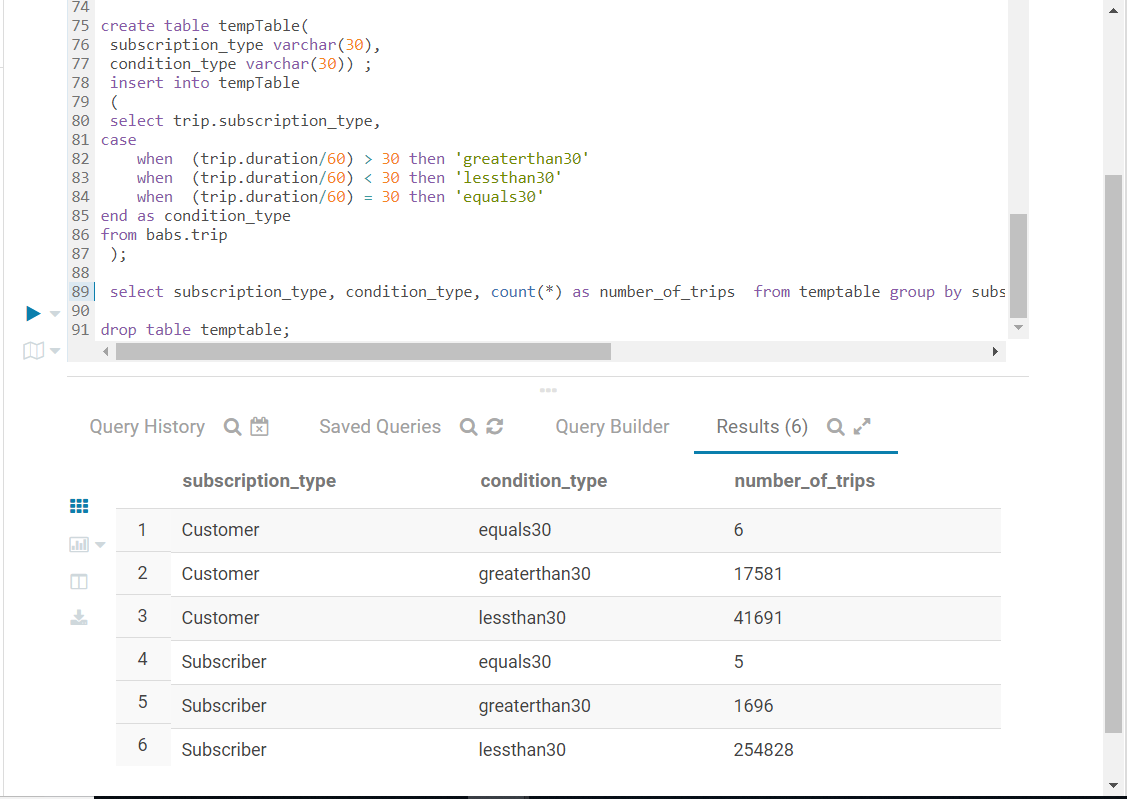
from babs.trip

);

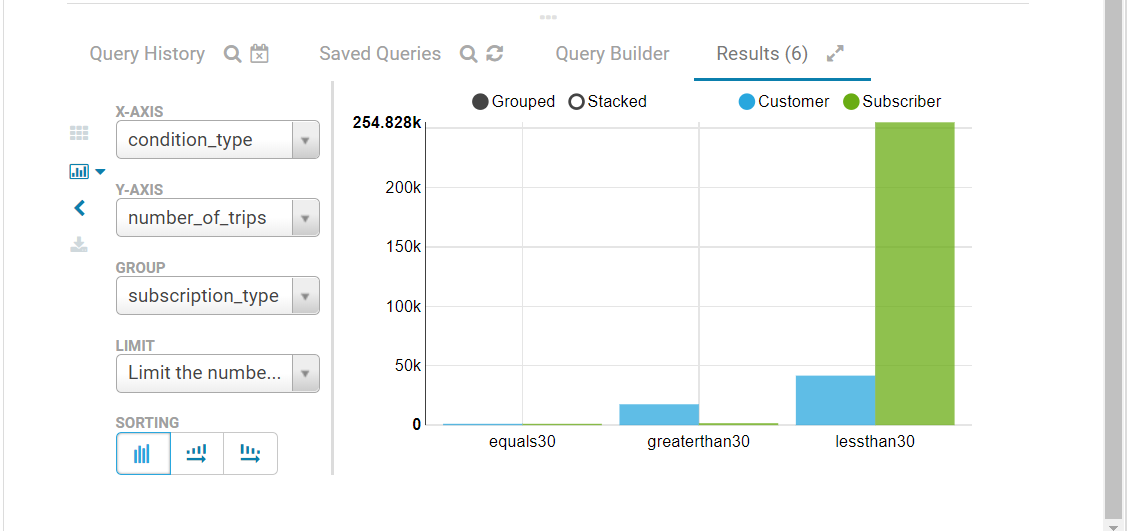
select subscription\_type, condition\_type, count(\*) as number\_of\_trips from temptable group by subscription\_type, condition\_type;

drop table temptable;

**Output:**



**Visualization:**



**Explanation:**

Customers have made around 17581 trips which are greater than 30 minutes but subscribers have made only 1696 trips which are greater than 30 minutes. From this we conclude that subscribers are using intend way.

Subscribers have made 254828 trips which are less than 30 minutes but customers have made only 41691 trips which are less than 30 minutes.

So, we conclude that subscribers make many shorter trips compared to customers who do not pay monthly subscription

**12. Find the total count of trips grouped into Weekday and Weekend.**

**Query:**

create table tempTable(

trip\_id varchar(30),

start\_date varchar(30),

converted varchar(30),

daytype varchar(30)) ;

insert into tempTable (

select datepart.trip\_id, (from\_unixtime(unix\_timestamp(datepart.d , 'MM/dd/yyyy'), 'yyyy-MM-dd')) as start\_date,

(from\_unixtime(unix\_timestamp(datepart.d , 'MM/dd/yyyy'), 'EEEE')) as converted,

case

when (from\_unixtime(unix\_timestamp(datepart.d , 'MM/dd/yyyy'), 'EEEE')) = 'Monday' then 'Weekday'

when (from\_unixtime(unix\_timestamp(datepart.d , 'MM/dd/yyyy'), 'EEEE')) = 'Tuesday' then 'Weekday'

when (from\_unixtime(unix\_timestamp(datepart.d , 'MM/dd/yyyy'), 'EEEE')) = 'Wednesday' then 'Weekday'

when (from\_unixtime(unix\_timestamp(datepart.d , 'MM/dd/yyyy'), 'EEEE')) = 'Thursday' then 'Weekday'

when (from\_unixtime(unix\_timestamp(datepart.d , 'MM/dd/yyyy'), 'EEEE')) =

'Friday' then 'Weekday'

when (from\_unixtime(unix\_timestamp(datepart.d , 'MM/dd/yyyy'), 'EEEE')) = 'Saturday' then 'Weekend'

when (from\_unixtime(unix\_timestamp(datepart.d , 'MM/dd/yyyy'), 'EEEE')) =

'Sunday' then 'Weekend'

end as daytype

from (select trip.trip\_id, substring(trip.start\_time,0,9) as d from babs.trip) as datepart

);

select daytype, count(daytype) as no\_of\_trips\_byWeekDay\_WeekEnd, sum(count(daytype)) over() as total\_trips from tempTable group by daytype ;

**Output:**



**Visualization:**

