Assignment 3 499100

1- Download the dataset from http://18.206.158.228/dataset/used_cars.csv into your home directory.

2- Using one line bash command print the number of columns.

```
[baoxin.l@ip-172-31-95-86 ~]$ head -1 used_cars.csv | tr "," "\n" | wc -l 23
```

3- Using the Linux commands, what are the three most frequent exterior colors? Did you expect these?

```
[baoxin.l@ip-172-31-95-86 ~]$ cut -d ',' -f 7 used_cars.csv | sort| uniq -c| sort -rn | head -3
1304 Black
840 White
481 Gray
```

- 4- Using the Linux commands, find the top 10 cars (combination of maker and model: like Ford
- F-150) that has "keyless entry". Remember that they may have written "keyless entry" with

either capital or small letters.

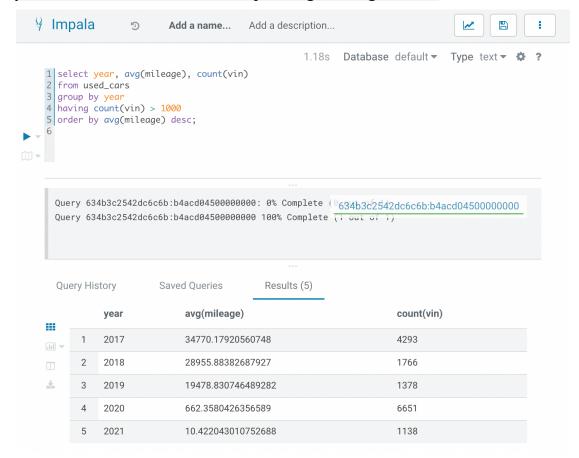
```
[baoxin.l@ip-172-31-95-86 ~]$ grep -i 'keyless entry' used_cars.csv | cut -d ',' -f 15,18 | sort | uniq -c | sort -rn | head -10 642 Jeep,Grand Cherokee 546 Ford,Fusion 457 Ford,F-150 340 Jeep,Compass 311 Ford,Escape 333 BMW,3 Series 295 Ford,Explorer 259 Chevrolet,Equinox 252 Hyundai,Tucson 245 Mazda,CX-5
```

5- Using the Linux commands, maximum mileage of Toyota SUVs. Make sure to keep track of words independent of capital/small letters.

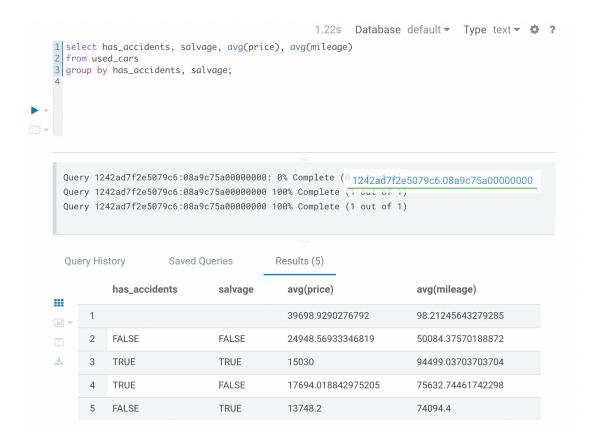
```
[baoxin.l@ip-172-31-95-86 ~]$ grep -i 'Toyota' used_cars.csv | grep -i 'SUV' | cut -d ',' -f 17 | sort -rn | head -10 261711 227763 221396 216954 215251 206996 204124 197644 197644 199548
```

6- Find the average mileage per years that where at least more than 1000 cars are listed in those

year. Sort the results to see the top average mileages first.



7- Find the average price and average mileage of cars per accident and salvage status.



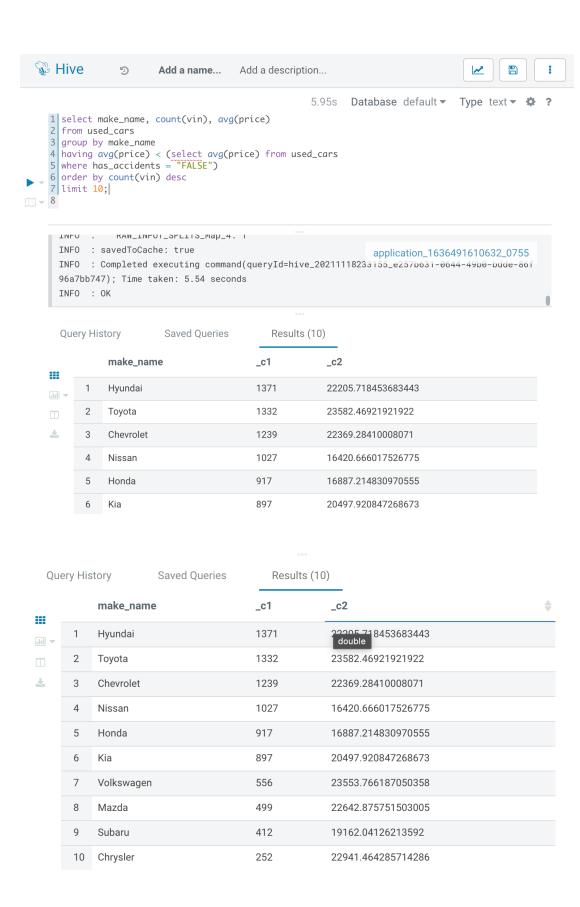
8- First the total car per car maker where the average price of cars by the car maker is less than

the average price of all cars without accident (accident as FALSE). Select the top 10 makes

with the most cars. Hint: You may not be able to run it on both of Impala and Hive. Try that

and let us know which one doesn't work.

Hive worked, Impala doesn't because it doesn't support having clause.



```
Impala

Add a name... Add a description...

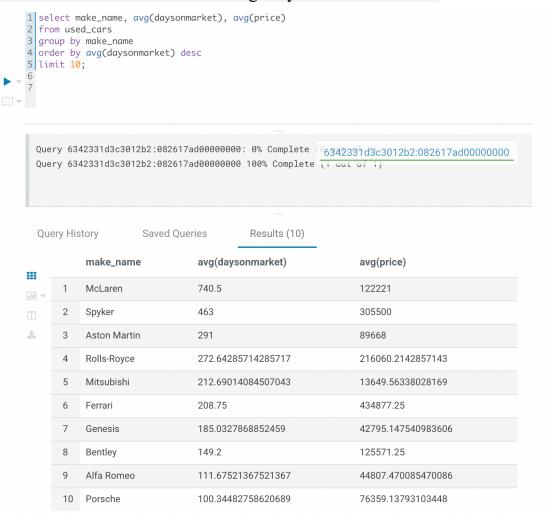
Os Database default ▼ Type text ▼ ?

1 select make_name, count(vin), avg(price)
2 from used_cars
3 group by make_name
4 having avg(price) < (select avg(price) from used_cars
5 where has_accidents = "FALSE")
order by count(vin) desc
1 imit 10;

AnalysisException: Subqueries are not supported in the HAVING clause.
```

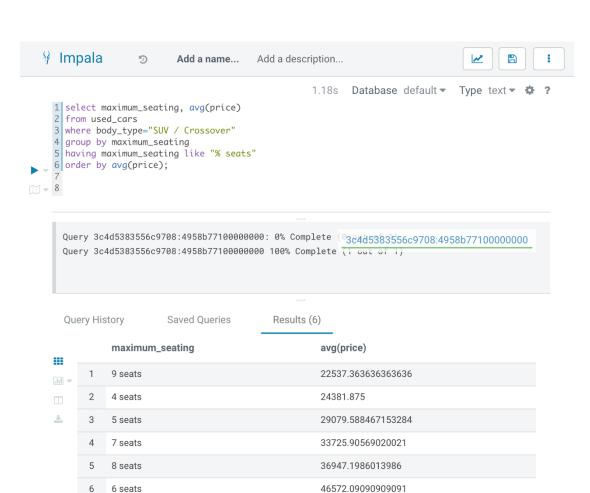
9- For each car maker find the average price and average days the car is on the market. List the

worst 10 brands based on the average days on the market.



10- Find the average price per maximum seat of SUV/Crossovers. List only valid seats (no null

or incorrect value). Sort the results by average price.



6 seats